In Silico and Wet Lab Approaches to Study Transcriptional Regulation

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Prof. mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 29 juni 2010
klokke 11:15 uur

door

Matthew Scott Hestand

geboren te Lexington, Kentucky, USA in 1979

Promotion committee

Promotores:	Prof.dr. G.J.B. van Ommen Prof.dr. J.T. den Dunnen
Co-promoter:	Dr. P.A.C. 't Hoen
Overige leden:	Dr. B. van Steensel (NKI) Prof.dr. P. de Knijff Dr. A.A.F. de Vries

The research described in this thesis was performed in the Department of Human Genetics, Leiden University Medical Center, The Netherlands. Chapter 3 research was in collaboration with the EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.

This work was supported by the Centre for Biomedical Genetics, the Netherlands, and the Centre for Medical Systems Biology, co-funded by the Netherlands Genome Initiative, and received additional funding by a Marie Curie Fellowship.

Printing of this thesis was supported by the Centre for Biomedical Genetics and the Netherlands Bioinformatics Centre (NBIC).

Printed by: Gildeprint Drukkerijen

ISBN: 978-94-6108-057-8

Contents

Chapter 1

Introduction

Today we have the technology to quickly sequence entire genomes, but annotating those sequences is still a daunting task. Discerning their function is even a more massive challenge. With only four nucleotides, the genome encodes tens of thousands of genes. Sequence content also determines regulation, providing sites for regulatory elements to control gene transcription. Regulatory elements that bind to genomic DNA can be in the form of proteins termed transcription factors (TFs). However, regulation goes beyond just sequence, encompassing epigenetic factors, from methylation to chromatin remodeling. To even further complicate the picture, regulation can occur at the RNA level by microRNAs, degradation, and alternative splicing. Translational control and post-translational modifications may also further determine the final gene product (a protein for many genes). The comprehensive picture is extremely complicated and too large for one individual to master. This thesis is devoted to one fraction of this picture: TFs and their target binding sites. We have studied two biological processes: the cell cycle (control) and myogenesis. By using a combination of *in silico* and wet lab work, including next-generation sequencing technology, we can better understand the TFs involved in transcriptional regulation of these processes, as outlined in this thesis.

1.1 Biological Background Information

Genetics and Genomics

The genomic code is embedded in our DNA, which is composed of a double helix of strands of nucleotides. There are four nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). DNA can be transcribed into RNA, composed of the same nucleotides other than T becoming uracil (U). RNA synthesis occurs in what is termed a $5'$ to $3'$ direction, using the DNA as a template. RNA in turn can be directly functional or translated into amino acids, the building blocks of proteins. Genetics is the study of genes. Classically genes were considered the portions of DNA that are transcribed into RNA, which is spliced in higher organisms. The portions of RNA (and corrosponding original DNA template sequence) that is retained after splicing are called exons and the portions removed are called introns. Most of the

spliced RNA is then translated into amino acids. What is not translated is called the untranslated region (UTR).

All the nucleotides in a person's DNA make up their genome. Traditionally, focus was only on all the genes in an organism. However, as our knowledge expanded to comprise regulatory elements not within genes the full genome became an interest of study. Genomics is the study of the genome. This includes how much of a gene is transcribed into RNA, termed gene expression, or transcriptomics. The number of genes in the human genome is difficult to know for sure. One explanation is that in one annotation transcripts may overlap constituting one gene and in another annotation the overlap may not be found indicating two separate genes. The initial sequencing of the human genome estimated 30,000 to 40,000 protein-coding genes (1) and a year later the number of genes was estimated to be closer to the low end of 30,000 protein-coding genes (2). Currently, as annotated by Ensembl v53 (3), there are 37,435 total genes (Biomart query, including non-protein coding genes (3; 4)). As more and more high-throughput datasets become available this number should become more reliable.

TFs and Promoters

TFs are regulatory proteins, or protein complexes, that bind to DNA, and positively or negatively influence gene expression. Pattern finding algorithms have been developed to identify TF binding sites (TFBSs) that are presumed to occur in a group of nucleotide sequences. A group of target nucleotide sequences could be known promoters. Promoters are typically a variable amount of base pairs (bp) surrounding the transcription start site (TSS) at the $5'$ end of a gene. For promoters, pattern finding is based on the presumption that promoters with similar regulation/expression have common regulators, and therefore similar TFBSs in their sequences. These regulating TFBSs should therefore have a high occurrence in similarly regulated/expressed genes' promoters.

TFs may also bind other TFs, and are then termed coactivators. There is evidence that some TFs may preferentially bind one strand of the DNA (5). Traditionally, the binding sites of TFs were looked for in the promoter region. One early example of promoter binding is Sp1, which binds the promoter region of beta globin genes (6), as well as 1641 promoters in an additional study(7).

Properly defining the promoter region of genes has been difficult. The promoter is often considered around the TSS, so the first exons of currently annotated genes indicate potential promoters. Many traditional annotation approaches have been results of sequencing RNA and aligning those sequences to a reference genome to infer exon locations. This was often done from the $3'$ end and the process was often considered complete when a full coding sequence was determined. Therefore, many exons with non-protein coding sequence were not annotated. In addition, genes do not necessarily have a single transcript per gene. Often, genes have multiple transcripts, comprised of different combinations of exons. This is a process that contributes to cells being different in one tissue than another. Due to these alternative transcripts, the promoter being used in a specific cell may be around a different exon than the annotated first exon of a gene.

Promoters may also be divided into several classes. Some promoters, such as those containing a TATA box (the target sequence of the polymerase II complex),

have TSSs with very specific locations, whereas others may contain broad TSSs with multiple positions of transcription initiation (8). The first class of promoters tend to be tissue specific, whereas the second class is more likely to be associated with house keeping genes (8). The latter promoters also tend to contain a higher number of GC dinucleotides than expected, termed CpG islands (8; 9; 10). CpG island promoters encompass a majority of mammalian promoters, whereas a minority of promoters are CpG poor (8). These issues are important to keep in mind since TFs may have a preference for one promoter type over another.

However, TFs may bind regions other than the promoter. When looking at some TFs, such as p53, TFBSs may also be located in introns and 3′ regions (11). TFs may actually bind far from a gene. These regions, which also regulate gene expression, are termed enhancers. The difference between promoters and enhancers is that both are regulatory regions, but promoters also contains the sites that basic transcriptional machinery binds to.

Whether a TF can bind its target DNA or not can also be regulated by the accessibility of the DNA. Open chromatin, accessible to the transcriptional machinery and associated with active gene expression, is termed euchromatin. Many epigenetic factors (not encoded by the DNA) are associated with euchromatin, including hypomethylation of CpG islands, multiple histone modifications and variants, and chromatin remodeling complexes (12). All of these factors can therefore have an influence on whether a TF can bind its target DNA or not. Whole regions of the chromosome, potentially containing multiple genes, may be regulated by what are termed locus control regions. One example is that of the locus control region for beta globin genes where binding of proteins to the locus control region play a critical role in multiple (up to 80 kb away) genes' activation (reviewed in (13)).

Better understanding TFs will give us greater knowledge into how the genome is regulated. In a larger view it may help us to even define what makes us human. With the high concordance between coding DNA in the human and chimpanzee (>99% at the protein level) it has long been believed that what largely makes us human is not the genes themselves, but the regulation of their transcriptome (14).

The Cell Cycle

One of the hallmarks of living cells is the process of cell duplication. This so-called cell (division) cycle is a tightly regulated process due to the expression and activation of stage-specific proteins that control the different cell cycle transitions (G1/S, S, and G2/M phases; reviewed by Satyanarayana and Kaldis, 2009 (15) and Malumbres and Barbacid, 2009 (16)). Loss of control of the cell cycle can lead to increased cell proliferation, resulting in tumors. By better understanding the regulators of the cell cycle scientists hope to guide research into cures for diseases such as cancer. Chapter 5 of this thesis involves a study of TFs which play a role in the cell cycle.

Many factors contribute to cell cycle regulation, including hormones, growth factors, cytokines, cyclin-dependent protein kinases, cyclins, the retinoblastoma (RB) protein, bcl-2 protein, myc protein, bax protein, the E2F family of TFs, and the TF p53 (17; 18). The tumor suppressor p53 is a crucial cell cycle regulator, with an estimated 50% of tumors carrying a mutation in the p53 encoding gene (18). In Chapter 3, based on *in silico* predictions, we identify TFs that potentially cooperate with p53.

p53 itself can be regulated by coactivators such as p300 and CBP (19). Besides by interactions with TFs, these acetyltransferases also regulate gene expression by altering chromatin accessibility via the acetylation of proximal nucleosomal histones. Despite their high levels of homology, the coactivators are not able to substitute for each other during embryogenesis as was shown by mouse knockout experiments (20; 21). Thus, in chapter 5 we selected these two coactivators for study.

Myogenesis

Several other parts of this thesis (chapters 2, 3,and 6) aim at elucidating the roles of TFs regulating myogenesis. Myogenesis is the process of muscle formation and development. The process of myogenesis may be divided into two parts: embryonic and adult. During embryogenesis somites develop into mesodermal precursor cells (22; 23). These mesodermal precursor cells are pushed towards a myogenic lineage by two primary myogenic TFs: MyoD and Myf5 (22). These resulting cells are termed myoblasts, which further differentiate into primary and secondary myofibers.

We focus on the process of adult myogenesis, through which myoblasts cease proliferating and fuse together to form multinucleated myofibers. In skeletal muscle this was traditionally and simplistically believed to be controlled by four major TFs: MyoD, Myf5, Myogenin, and MRF4, with the first two functioning in early differentiation and latter two in late differentiation (24). However, as our knowledge of biological pathways and processes expands it is becoming apparent that many TFs and other elements are responsible for the regulation of myogenesis. Besides the four major TFs, Charge *et al.* 2004 review many molecules, including other TFs (including Pax7, Pax3, Slug, myocyte nuclear factor (MNF), and Msx1) that contribute to myogenesis (22). A year later an initial blueprint of myogenic differentiation was published including MyoD, Myogenin, and MEF2 targeting a large number of additional TFs, with connections being made to TEAD4/TEF-3, ARNT, Copeb/KLF6, NFE2l2/NRF2, and ATF4 (25). As genetics moves forward it is likely more and more TFs will be identified that play a role in myogenesis.

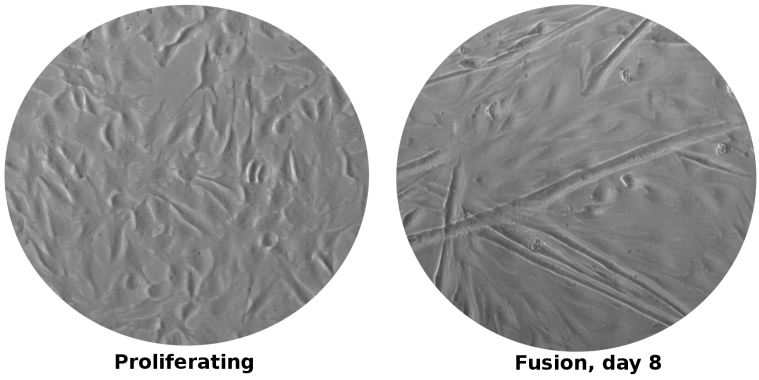


Figure 1.1: Proliferating and Differentiating Mouse C2C12 cells

Several systems exist to study myogenesis in the laboratory. These includes patient

samples, mouse strains, and cell lines. This thesis primarily uses a mouse cell line termed C2C12. These cells proliferate with serum, but when serum deprived stop proliferating and begin to differentiate and fuse into myotubes (Figure 1.1). This process typically takes seven to nine days.

Defects in myogenic regulation (via a TF mutation or alteration of its target) result in a multitude of diseases, including myotonic dystrophy, rhabdomyosarcomas, Waardenburg syndrome type 2, congenital myasthenia, and diseases related to muscle regeneration (overview in Martin 2003 (26)). By gaining a better understanding of the genetic architecture of late myogenesis we hope to aid researchers towards developing cures for such illnesses.

1.2 Conventional Wet-lab Methods

RNA Expression

Many kits and techniques now exist for the isolation of RNA. A traditional method for over twenty years is an extraction with guanidinium thiocyanate, Phenol-chloroform, and sodium acetate, followed by isoproponal precipitation clean-up (27).

Serial Analysis of Gene Expression (SAGE) (28) (Figure 1.2) and Cap Analysis of Gene Expression (CAGE) (29) (Figure 1.3) are two methods to isolate small parts at either end of mRNAs. These were classically concatenated and cloned into libraries and then sequenced. With next generation sequencing technology (see below) it is possible to directly sequence the SAGE/CAGE sequences (termed DeepSAGE (30) and DeepCAGE (31)).

SAGE is a method developed to quantify all the transcripts expressed in a genome (28). This commonly works by isolating RNA poly-A tails with oligo(dT) beads, converting into cDNA, performing a first restriction digest (NlaIII which cuts at CATG's), retaining the $3'$ most fragments, adding a linker to the $5'$ end with a restriction site, then using an additional enzyme that recogizes the linker site (such as MmeI) to cut a certain number of bp from the $5'$ end each fragment, typically 14-20, adding a second linker to the $3'$ end, and finally cloned and sequenced (32) (Figure 1.2).

CAGE is a technique to sequence the $5'$ end of transcripts and therefore better annotate TSSs, which can be used to provide better promoter annotation (29). CAGE works first by creating single strand cDNA and then capturing the $5'$ cap, present on all mRNAs, with an antibody or biotinylated cap-trapper (Figure 1.3)(29). A linker sequence is then added to the $5'$ end which contains sequence to bind to the sequencer's glass slide (for Illumina next-generation sequencing), a sequencing primer, and a restriction enzyme site. Double strand cDNA synthesis is then performed and a restriction enzyme actually cuts a number of bp downstream of the linker restriction enzyme site, providing approximately 20-26 bp of the original $5'$ end of the transcript. A final linker is added for the sequencing protocol and in current protocols the library is run through a next-generation sequencing machine.

In contrast to $5'$ or $3'$-end focused methods, true whole transcriptome sequencing, also called mRNA-seq, is a method by which cDNA generated on the total RNA by random priming is amplified, sheared, and sequenced (33). This method therefore provides a more complete picture of RNAs, but can be more complicated for expression

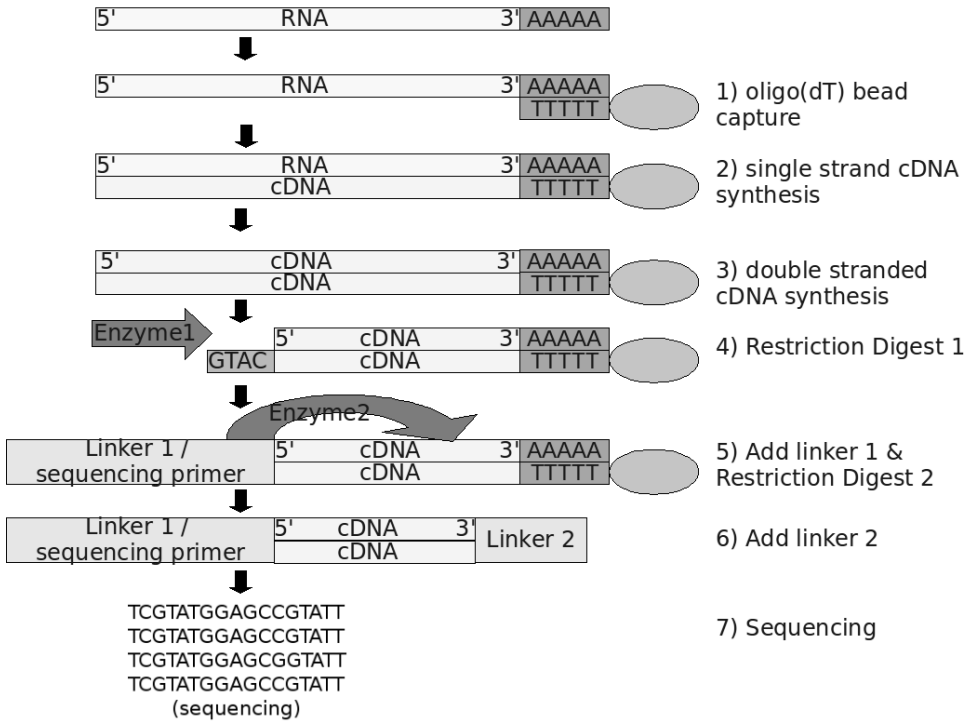


Figure 1.2: DeepSAGE: Sequencing of serial analysis of gene expression libraries starts by capturing the poly-A tail of mRNAs with oligo-dT beads. RNA is converted to cDNA and made double stranded, followed by a first restriction digest. The 3′ most fragments are retained, sequencing specific linkers adapted with a restriction site, and a second restriction digest performed that cuts downstream of the introduced restriction site. A second linker sequence is adapted and next-generation sequencing can then be performed.

analysis since one transcript may be represented by a greater diversity of tags. Having multiple random tags per transcript also reduces the quantity of total transcripts detected, reducing statistical power for calling differential expression levels. This is increasingly offset by the major increases in sequencing depth.

Isolating TF bound DNA

Chromatin immunoprecipitation (ChIP), is a wet lab technique to identify the targets of a specific TF (Figure 1.4). In general, this technique begins by formaldehyde fixing cells so that the TFs are fixed to the DNA. The cells and nucleus are then lysed, often with detergents, and the chromatin (DNA bound by RNA and protein) is isolated and cleaned up. This chromatin is then fragmented with chemicals or sonication. TF bound fragments of chromatin are then immunoprecipitated using an antibody targeting the TF of choice. This isolated pool of TF bound chromatin fragments

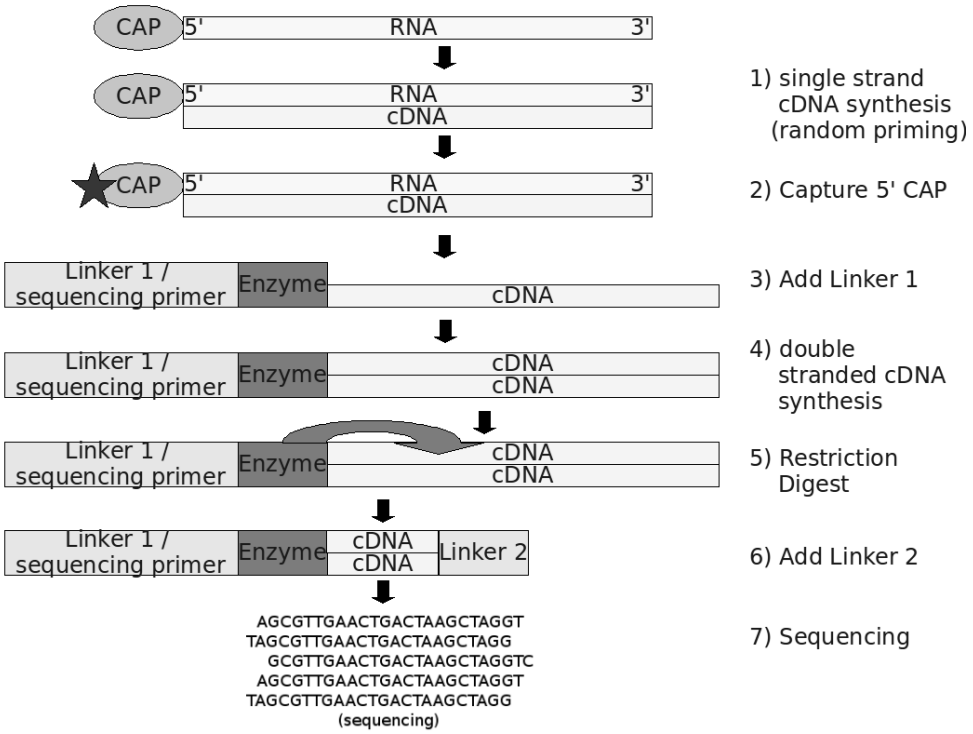


Figure 1.3: DeepCAGE: Sequencing of cap analysis of gene expression libraries starts with random priming for single strand cDNA synthesis and then capturing RNAs by their 5′ cap. A linker with a restriction site and sequencing linker is ligated and double stranded cDNA synthesized. A restriction enzyme is used that cuts downstream of the restriction site. The 5′ fragments are retained and a second linker ligated to the 3′ end of the fragment. These linker adapted sequences can then be applied to next-generation sequencers.

are then reverse cross-linked and cleaned up to leave only DNA fragments that were originally bound by the TF of interest.

The ChIP wet-lab method can be coupled with several genomic technologies to analyze ChIP target sequences genome-wide. When ChIP sequences are hybridized to a microarray (see below) it is termed ChIP-chip (or ChIP-on-chip) (34). An alternate approach is massive parallel sequencing, either with a paired-end ditag approach (ChIP-PET) (11), or directly using a next-generation sequencer (ChIP-seq) (35), as addressed below. These methods both start with ChIP, resulting in a pool of TF bound DNA. In ChIP-PET these are cloned into a plasmid vector, converted to concatenated and cloned PETS, and then sequenced (11). ChIP-seq is less laborious, omitting the cloning and concatenation steps, by just directly ligating linkers and sequencing the ChIP DNA.

Only several ChIP-seq experiments have been published at the time of this thesis, though large numbers of ChIP-chip studies have been published. ChIP-seq is expected

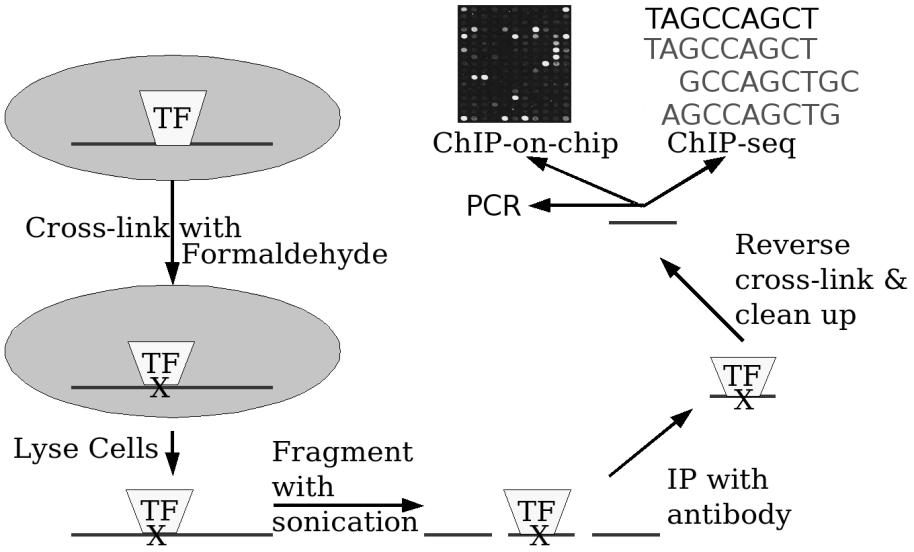


Figure 1.4: ChIP techniques: Chromatin immunoprecipitation (ChIP) works by cross-linking TFs to the DNA with formaldehyde, lysing cells, fragmenting chromatin with sonication, immunoprecipitating TF bound DNA fragments with an antibody, reverse cross-linking to remove TFs, and cleaning up the final pool of DNA fragments (originally bound by the TF of interest). These pools of DNA fragments can be analyzed by PCR, microarray (ChIP-(on)-chip), or next-generation sequencers (ChIP-seq).

to be an up and coming technology. It has the advantage, in comparison to ChIP-chip, of requiring less input material, the potential to identify TFBSs with low affinity, not being limited to target regions (*i.e.* probes on a microarray), not having hybridization errors, and is less costly for whole genome analysis (35). This method will rewrite the books on how TFs bind genome wide, identifying many TFBSs in intragenic regions that were not studied previously or were bound at too low a concentration to be detected by microarrays. This additional wealth of data will provide more sequences to mine for position weight matrices (PWMs, see below) and improve upon existing PWMs, resulting in improved *in silico* predictions.

Single Target Readout methods

The polymerase chain reaction (PCR) is a method to amplify a stretch (up to several kilobases) of DNA. DNA regions are targeted using primers specific to the DNA region. A polymerase is used to read the DNA and replicate it. The simplest use of this is to see if the DNA stretch is present in the genome. After amplification the product can be viewed on an agarose gel, and if appropriate markers are included the size can be estimated. The intensity of a band (compared with a control sample) on this gel can represent the relative quantity of DNA in the original sample, but to be more precise an adaptation of PCR is used. Quantitative real-time PCR, termed qPCR, uses fluorescent dyes or probes to quantify the amount of target DNA. RNA

can also be converted to cDNA with a reverse transcriptase and qPCR performed, termed RT-qPCR. This is especially useful and cost-effective to determine expression levels of RNA because of simplicity and high sensitivity.

Other methods also exist to detect TF bound DNA. This includes luciferase assays, deletion constructs, gel shift assays, and the TransFactor kit. Luciferase assays are a technology in which a promoter from a gene is cloned in front of a gene encoding a luciferase gene. When activating TFs bind to this promoter they activate the luciferase gene, causing the cell or organism to produce light under proper conditions. Deletion constructs are a means of eliminating a portion of a gene's promoter, then observing the effect.

Gel shift assays, involves running DNA through a gel. If a stretch of DNA has a TF bound to it, the sequence will run out slower on a gel. This is a relatively faster method than the previous two, but only indicates binding and no regulatory function. The TransFactor kit works on a similar level, determining binding of a TF to a target DNA sequence using a TF specific antibody, a secondary antibody, and colorimetry.

High-throughput Readout methods

Microarrays were one of the first technologies to study genetics at a genomic scale in a single test. Microarrays traditionally consist of a glass slide with thousands or millions of probes attached to it. These probes have sequences that bind target sequences. The target sequences are labeled with a dye that cause bound probes to give a fluorescent signal. Therefore, spots, consisting of clusters of probes, give a signal relative to the quantity of their target sequences in the sample analyzed. The most common use of microarrays involves hybridizing RNA to study gene expression levels.

Alternatively, microarrays used in conjunction with ChIP can search for a large number of TF targets. Promoter based and whole genome tiling arrays also exist to analyze the afore mentioned ChIP samples. These arrays consist of probes that are "tiled" (spaced) across promoters, or the entire genome. These can provide ideal target regions to study ChIP.

In the past few years several new technology platforms have emerged that perform DNA sequencing on a massive scale at a fraction of the speed and cost of traditional sequencing technologies. The primary three systems are the 454 by Roche, the Illumina Genome Analyzer (formerly Solexa) by Illumina, and the SOLiD by Applied Biosystems. Our department has two Illumina Genome Analyzers so this thesis's next-generation sequencing (NGS) has been performed on this system.

Though all classified as second or next-generation sequencers, these platforms have very different mechanics. The 454 is based on attaching DNA fragments to beads (one fragment to one bead), emulsion PCR amplification of the fragments on the beads, and loaded onto a PicoTiterPlate (one bead per well) for sequencing (*www.454.com*). Sequencing is performed by sequentially adding complementary nucleotides that emit a florescent signal, detected by a camera (*www.454.com*). SOLiD also uses beads and emulsion PCR, but then the amplified products are applied to a glass slide (*www.appliedbiosystems.com*). Several series of ligations are performed in which fluorescently labeled di-base probes are used for detection (*www.appliedbiosystems.com*). This system differs in that a fluorescent signal does not reflect the addition of an exact

nucleotide, but a pair (which is termed colorspace) (*www.appliedbiosystems.com*). Illumina differs in that no beads or emulsion PCR are used. Adapter ligated sequences are first attached to a slide, and then bridge amplification is performed on the slide (*www.illumina.com*). Nucleotides are then sequentially added which emit a different fluorescent signal for each of the four nucleotides, which is recorded by a camera (*www.illumina.com*).

Table 1.1: Next-Generation Sequencing System Specifications

Company	Applied Biosystems	Roche	Illumina	Applied Biosystems
Machine	traditional sequencing (3730xl DNA Analyzer)	FLX Titanium	Genome Analyzer IIx	SOLiD 3 Plus System
read length	up to 900 bp	400-500* bp	35-100 bp	35-100* bp
# reads per run	96 or 384 x 16 plates	∼1 million	∼150-200 million*	∼200 million*
run time	0.5-3 hours	10 hours	2-9.5 days**	3.5-14 days**
reference	*www.applied biosystems.com*	*www.454.com*	*www.illumina.com*	*www.applied biosystems.com*

*Numbers from website adapted based on personal experience. **Run times depend on the number of cycles (bp sequenced per read). Machine details are based on website specifications in February 2010.

These systems can produce vast amounts of data, however the read length, total bp sequenced, and sequence time vary between instruments (Table 1.1). The read length and total bp sequenced are also continuously increasing with advancements in chemistry and mechanics. It has been shown that next-generation sequencers outperform microarrays in precision, reproducibility, and sensitivity, likely by avoiding the problems associated with hybridization techniques (36). NGS (also called deep-sequencing or second-generation sequencing) also escapes the limitation of only looking at the targets that have been spotted on a microarray, *i.e.* performing a "content-limited" analysis.

Typically NGS analysis begins by converting data to sequences and filtering for quality. For the Illumina Genome Analyzer, this means converting image files and filtering on quality with their pipeline. For most NGS applications the next step is to align to a reference genome. Traditionally for longer reads alignments could be done with BLAST (37) or BLAT (38), but these algorithms do not perform well with large numbers of short reads, such as those provided by the Illumina Genome Analyzer. To align short reads many different alignment algorithms have been developed in the past years, including Eland (part of the Illumina GA Analysis Pipeline: fast, but only good for reads $\leq$ 32bp), Maq (39), Rmap (slow, but accurate) (40), Cloudburst (fast and accurate, but large system requirements) (41), Bowtie (fast) (42), and BWA (fast) (43). When a reference genome is not available, sequences are often built into contigs with the tool Velvet (44). From here analysis is very dependent on the application being analyzed.

1.3 *In silico* Prediction of TFs and TFBSs

Pattern Finders

As mentioned earlier, pattern finding algorithms can be used to identify TFBSs in sets of TF bound DNA sequences. Modern pattern finders include MEME (45; 46) and Gibbs samplers (47; 48; 49), which can find one or more variable patterns in DNA or protein sequences.

Position Weight Matrices

One method to identify TFBSs for known TFs is using PWMs (50). These matrices summarize experimental information on the sequential preference of a TF (Figure 1.5). The two leading databases of experimentally determined PWMs are TRANSFAC (51; 52) and JASPAR (53; 54). TRANSFAC has the advantage of more PWMs (834 matrices (release 11.4, December 2007)) (52) compared to JASPAR (123 matrices) (54). However, to use the larger TRANSFAC Professional (there is also a smaller public version free to all non-commercial users) a paid license is required, whereas JASPAR is free. These PWMs are used by programs like Match (51; 55) or Sunflower (56) to identify TFBSs in a nucleotide sequence by evaluating the nucleotide similarity of the PWM with the sequence.

	C	A	T	G
Nucleotide 1:	0	0	10	0
Nucleotide 2:	1	4	4	1
Nucleotide 3:	0	0	5	5
Nucleotide 4:	8	0	1	1
Nucleotide 5:	1	1	7	1

Figure 1.5: A Theoretical Position Weight Matrix (PWM): At the top is a theoretical chart of a 5 nucleotide PWM made up from 10 experiments. For each nucleotide is a count of how many experiments found that nucleotide. Below is shown a visual representation of the chart information.

Over-Representation of TFBSs

However, even with PWMs, identifying TFBSs is a difficult task, considering genomes may be in the billions of base pairs and TFBSs may be only 12-14 bp in size (49).

One method to improve upon TFBS predictions in a set of genes is to look for over-representation of TFBSs in the promoters of co-regulated/co-expressed genes. Using a similar presumption as described for pattern finders, it is presumed that

similarly regulated/expressed genes' promoters contain common regulators. Therefore, target TFBSs identified through PWMs should occur more often in a similarly regulated/expressed set of genes' promoters than in a random set of genes' promoters. This method has been developed to include work on complex organisms such as human (57). This method relies on using proper target sequences. Therefore, good gene/promoter annotation is critical, such as that provided by CAGE techniques.

Conservation of TFBSs

Another method to look for *de novo* TFBSs is by searching for conservation between orthologous promoters (58). This method is based on the presumption that functional elements are evolutionarily conserved and mutations in these elements could therefore be detrimental to the organism (58; 59). Programs that use conservation to determine TFBSs include oPOSSUM (60) and ConTra (61).

1.4 Thesis Overview

This thesis looks at TFs and TFBSs discovery first through *in silico* predictions based on previous ChIP and expression data, then wet lab work with *in silico* confirmation. Chapter two focuses on CORE_TF, a web site developed to identify over-represented and cross-species conserved TFBSs in a set of similarly regulated genomic regions, such as up-regulated genes' promoters from a microarray study. The third chapter achieves a similar goal to chapter two to identify over-represented TFBSs, but also models competition between TFs, which better models the true biological system and, thus, improves results. Chapter four presents a pipeline, titled GAPSS, to analyze NGS data that was used for data analysis of chapters five and six. Chapter five focuses on ChIP-seq wet-lab work and data-analysis, including GAPSS and CORE_TF, to better understand the role of CBP and p300 in cell cycle control. The sixth chapter primarily focuses on using CAGE to better annotate muscle specific TSSs which should improve promoter based TFBS predictions. Chapters seven to nine wrap up this work, explaining how a combination of multiple *in silico* and wet lab techniques lead to a better understanding of the transcriptional control of genes.

Chapter 2

CORE_TF: a User-Friendly Interface to Identify Evolutionary Conserved Transcription Factor Binding Sites in Sets of Co-Regulated Genes

Matthew S. Hestand, Michiel van Galen, Michel P. Villerius,
Gert-Jan B. van Ommen, Johan T. den Dunnen, Peter A.C. 't Hoen

The Center for Human and Clinical Genetics, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.

BMC Bioinformatics 2008, 9:495
Parts of this manuscript have been adapted to more appropriately fit this thesis.

2.1 Abstract

Background: The identification of transcription factor binding sites is difficult since they are only a small number of nucleotides in size, resulting in large numbers of false positives and false negatives in current approaches. Computational methods to reduce false positives are to look for over-representation of transcription factor binding sites in a set of similarly regulated promoters or to look for conservation in orthologous promoter alignments.

Results: We have developed a novel tool, "CORE_TF" (Conserved and Over-REpresented Transcription Factor binding sites) that identifies common transcription factor binding sites in promoters of co-regulated genes. To improve upon existing binding site predictions, the tool searches for position weight matrices from the TRANSFAC^R database that are over-represented in an experimental set compared to a random set of promoters and identifies cross-species conservation of the predicted transcription factor binding sites. The algorithm has been evaluated with expression and chromatin-immunoprecipitation on microarray data. We also implement and demonstrate the importance of matching the random set of promoters to the experimental promoters by GC content, which is a unique feature of our tool.

Conclusion: The program CORE_TF is accessible in a user friendly web interface at *http://www.LGTC.nl/CORE_TF*. It provides a table of over-represented transcription factor binding sites in the users input genes' promoters and a graphical view of evolutionary conserved transcription factor binding sites. In our test data sets it successfully predicts target transcription factors and their binding sites.

2.2 Background

There are both experimental and computational approaches to identify transcription factors (TFs) and their relevant binding sites. In the wet lab, hypothesis driven techniques, such as deletion constructs with luciferase reporter assays and chromatin-immunoprecipitation on microarrays (ChIP-on-chip), can be used to identify TF binding site (TFBS) regions. Luciferase assays can prove that a specific region has regulatory function, but they are laborious and time consuming. ChIP-on-chip is more global, but requires prior knowledge of which TF to target using a specific antibody and is laborious, time consuming, and expensive. Faster and cheaper *in silico* methods have been in development which can identify potential TFs and their binding sites. They also tend to target more precise the TFBS instead of just containing a TFBS region. However, finding TFBSs can be extremely difficult since they may be less than 12-14 bp long and their consensus binding sites may be fairly loose (49).

One method to identify TFBSs for known TFs is using position weight matrices (PWMs) (50). PWMs summarize experimental information on the sequence preference of TFs. TRANSFAC (51; 52) is the leading PWM database for TFBSs with 834 matrices in total (release 11.4, December 2007), compared to 123 in JASPAR (53; 54).

An additional method to look for new (*de novo*) TFBSs is by searching for conservation between orthologous promoters (58). This is based on the presumption that functional elements are evolutionary conserved since mutations to such elements could be detrimental to the organism (58; 59).

However, both the sequence conservation-based and the PWM approach alone produce many false positives and false negatives. We therefore created CORE_TF, a program using both methods to reduce false predictions. We first look for TFs involved in a biological process of interest, relying on the presumption that similarly expressed genes have common TFs as regulators. To do this, and reduce false predictions with PWMs, we search for TFBSs that occur more often in a co-regulated set of promoters compared to random promoters. This algorithm, in analogy to the work of *Elkon et al*, 2003 (57), implements a binomial test to evaluate for this over-representation. Some PWMs have a bias towards certain nucleotides, such as T's and A's for a TATA box binding TF and would therefore likely be over-represented if an experimental set had high numbers of T's and A's and the random set had equal content of all four nucleotides. We therefore also offer the option to exclude biases based on GC content by matching random promoters with approximately equal GC content to the experimental promoters. To identify individual TFBSs with increased precision, and add additional support for the relevant TFs, we subsequently scan individual promoters for cross-species conservation, again employing TRANSFAC matrices. All steps are flexible allowing for a multitude of input types (Ensembl (62) gene IDs, nucleotide sequences, or selected by CORE_TF).

We also compared CORE_TF to two existing programs: oPOSSUM (60) and ConTra (61).

CORE_TF is accessible as a web-page. In this paper, we present and evaluate the performance of our web-based tool for identification of TFBSs.

2.3 Implementation

2.3.1 CORE_TF Construction Format

The main script is written in Perl and presented in HTML on an Apache web-server. Input and table sorting is done using an edited Java script: sorttable.js (63). By default, following the title page, there are 6 pages that are run in a linear fashion feeding the results of one page into the next (Figure 2.1).

Page one allows a user to select run options and input criteria, including a p-value cut-off for highlighting data (see below), 6 different Match (the program that aligns TRANSFAC PWMs to nucleotide sequences) (51; 55) settings (minimize false positives, minimize false negatives, minimize the sum of both error rates, and non-redundant sets of these 3 settings), and data input type for a set of experimental promoters and a set of random promoters. The experimental promoter lists are entered as sequences in fasta format or Ensembl gene IDs. Five options are available for the random promoter list input: sequences in fasta format, an Ensembl gene ID list, randomly retrieve Ensembl promoters, pre-constructed promoter sets, and pre-retrieved sequence sets that are matched to the experimental set based on percentage of GC content. There is also an option to skip the over-representation analysis and go directly to page 4.

Depending on the selections from page 1, page 2 presents text boxes to paste in lists of fasta format sequences or Ensembl gene IDs, or radio-buttons to select a certain number of random promoters for the appropriate species, or species based check boxes for pre-constructed runs or %GC matched runs. If CORE_TF must retrieve promoters there are two options to define promoter sequences. The first option is to call a promoter as exon 1 plus a user defined number of base-pairs (bp) upstream. The second option is to define a promoter sequence as a user specified number of bp before and after the start of exon 1. The pre-constructed (approximately 3000 promoters) and pre-retrieved sets to match %GC on (approximately 10000 promoters, of which 3000 are selected) are based on 1000 bp upstream of exon 1 and exon 1 sequence.

If requested, page 3 (Figure 2.2) uses Ensembl API to retrieve promoters from a locally installed Ensembl database or from the web-based Ensembl database depending on CORE_TF installation. If the option to use %GC matched random sequences is selected CORE_TF matches pre-retrieved promoter sequences to the experimental promoter sequences so that at least 3000 similar %GC promoters are obtained. It then uses Match to scan all sequences for the presence of TRANSFAC Professional (note: web based CORE_TF is still free access to non-commercial users) vertebrate PWMs passing the PWMs' alignment threshold provided on page 1 (pre-constructed random promoter sets also have pre-executed Match runs and initial number of hits counted). A binomial test is carried out with the Perl module Math::Cephes (64) to identify TFBSs that are over-represented in the experimental set over the random set. This is displayed on the screen as a sortable table with the TFBSs' name, p-value (10 digits are displayed), hits and total number in the experimental and random sets, as well as the number of PWM hits in each experimental promoter. For clarity, p-values below a defined threshold from page 1 are highlighted in blue. The table can be downloaded as an HTML file or a tab-delimited text file. The user can select a number of TFBSs plus a promoter of interest and continue to the next page. There is also a Java script with a button to automatically select all TFBSs with a p-value

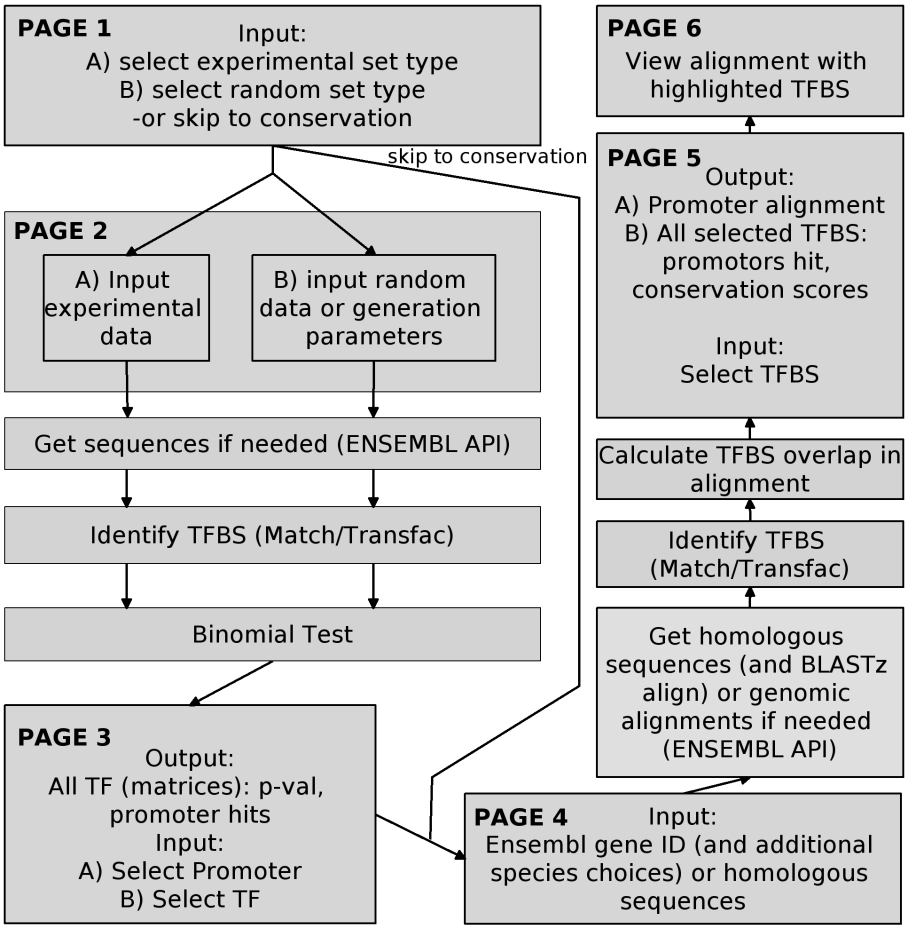


Figure 2.1: Flowchart of CORE_TF runs: CORE_TF runs linearly through 6 web pages. Pages 1 and 2 take as input experimental gene/promoter lists and random gene/promoter lists or requests to create random lists. Depending on format, sequences are retrieved with Ensembl API or random lists generated before identifying TFBSs with Match/TRANSFAC. A binomial test is run to identify over-represented TFBSs in the experimental set compared to the random set and displayed in page 3 as a table. In the table TFs and a promoter can be selected which are sent to page 4. If requested homologs and sequences or genomic alignments are retrieved from Ensembl for the selected promoter. If not already a genomic alignment, input sequences or retrieved sequences are aligned with BLASTz. TFBSs are identified with Match/TRANSFAC, overlapping TFBSs are identified and scores calculated, and the data is displayed in page 5. Conserved TFBSs can be selected and displayed as highlights in the alignment in page 6.

below the defined threshold.

Page 4 gives the user the opportunity to use Ensembl defined orthologs or aligned

TFBS output

Match cut-off was minimize the sum of both error rates
These are high quality vertebrate (V$) matrix

p-value high-light has cut-off below 0.05
NA = the frequency of the TFBS in the experimental or random promoters was 0 or 1

Select in the table below the TF and promoter of interest for the homology analysis

Automatically check p-values below 0.05

[X]	Name of Matrix	P-values	# exp promoters hit	# exp promoters	# random promoters hit	# random promoters	freq random	hits in ENSMUSG00000031077 Fadd
☐	ACAAT_B	0.800125773	8	147	225	2940	0.08	1
☐	AFP1_Q6	0.946858085	7	147	178	2940	0.06	0
☐	AHRARNT_01	0.000056480	585	147	1244	2940	0.42	1
☐	AHRARNT_02	0.221496710	9	81	147	1537	2940	0.52	1
☐	AHRHIF_Q6	0.000031852	5	101	147	1553	2940	0.53	2
☐	AHR_01	0.005052041	2	49	147	714	2940	0.24	0
☐	AHR_Q5	0.061120733	0	15	147	210	2940	0.07	0
☐	ZIC3_01	NA	147	147	2924	2940	0.99	9
☐	ZID_01	0.096687210	5	30	147	491	2940	0.17	0
☐	ZNF219_01	0.221264173	4	12	147	204	2940	0.07	0
☐	ZTA_Q2	0.731407189	4	21	147	487	2940	0.17	0

Automatically check p-values below 0.05

Save table as HTML (right click-> save as).

Save table as TabDelimTxt (right click-> save as).

Select TF/Promoter to use in homology analysis

Submit Reset

Figure 2.2: Page 3 screen-shot: Page 3 of CORE_TF displays the following columns: selection boxes for the next page's analysis, all TFBS PWMs with hits, the p-value, the number of experimental promoters hit, the number of experimental promoters analyzed, the number of random promoters hit, the number of random promoters analyzed, frequency of hits in the random data, as well as a column for each experimental promoter analyzed indicating the number of TFBSs hit in it. Our page is lengthy, so for display purposes in this figure we deleted the middle TFBSs as indicated by the large black bar. For a full color figure see *www.biomedcentral.com/1471-2105/9/495/figure/F2.*

genomic regions in a selection of species (currently *H. sapiens, P. troglodytes, M. musculus, R. norvegicus, B. taurus, C. familiaris,* and *G. gallus*) or enter user defined orthologous sequences in fasta format. There is also the option to define promoters as was done in page 2. If the user skipped over-representation analysis there is a list of TFBSs to chose from for analysis, otherwise CORE_TF uses TFBS selection from page 3.

This is given to page 5 which, if necessary, retrieves either orthologous IDs and sequences or aligned genomic regions with Ensembl API. Aligned genomic regions are pairwise alignments, but CORE_TF places them into a multi-species viewed align-

ment. Sequences are again scanned by Match and TRANSFAC. If Ensembl genome alignments were not used, the first sequence entered or the ID used for orthologous retrieval is used as the reference to carry out a promoter sequence alignment with BLASTz (65). Alignments are displayed on the screen. Tables are shown with each TFBS selected and the following information: total score, region score, number of promoters aligned at that point, and the length of the TFBS. The region score is defined by taking the sum of 100 times the percent of each nucleotide aligned (Figure 2.3A). The total score is defined as the region score divided by the pattern length divided by 100 (Figure 2.3B). More specific details of these region numbers are displayed on additional tables lower in the page. The user may select a TF and submit this to the final page.

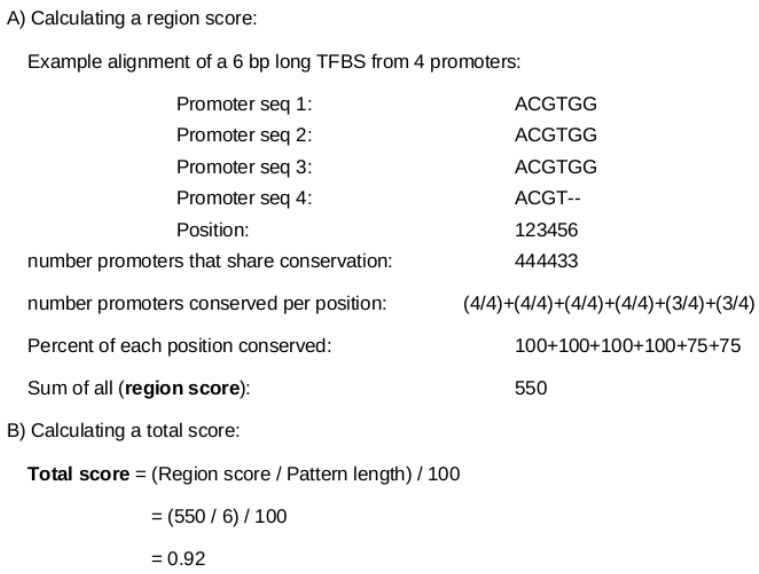


Figure 2.3: Formulas for conservation scores.

Page 6 (Figure 2.4) allows for visualization in the alignment by displaying the alignment with selected TFBSs highlighted according to the strand bound: blue (positive strand), purple (both strands), or red (negative strand). There is also evidence that some TFs may preferentially bind one strand over the other (5). It is up to the user to decide if their TF is strand specific or not.

2.3.2 CORE_TF Evaluation with Expression and ChIP-on-chip Data

To verify the performance of our algorithms we used expression and ChIP-on-chip data from *Cao et al* 2006 (66). They studied the promoter binding of two major regulators of muscle differentiation (MyoD and Myog) and expression profiles in embryonic fibroblasts from MyoD/Myf5 knockout mouse transduced with a MyoD-estrogen receptor hormone binding fusion protein (termed MDER cells). These cells have been modified so that they can be studied during differentiation with or without MyoD or

Figure 2.4: Page 6 screen-shot of a conserved MyoD TFBS in the *LAMA4* promoter: Page 6 of CORE_TF displays two identical boxes containing aligned promoters with conserved TFBSs highlighted by color; blue if on the positive strand, purple if on both strands, and red if on the negative strand. For a full color figure see *www.biomedcentral.com/1471-2105/9/495/figure/F4*. If requested in the previous page to show run details (not shown in this figure), boxes with score construction for all conserved TFBSs are also displayed, as well as the patterns of all selected PWMs hit. Here we show an example of a MyoD TFBS (PWM MyoD_Q6_01) in the *LAMA4* promoter conserved in human, chimp, and dog on both strands.

Myog present. Promoter binding was also studied in a common mouse myoblast cell line (C2C12).

ChIP-on-chip is a technique using a TF targeting antibody that is used to pull-down TF bound DNA fragments, which are then amplified, labeled, and hybridized to a (promoter or tiling) microarray. As a positive control set for TF binding, we took those promoters from the ChIP-on-chip data that showed enrichment for MyoD or Myog binding sites (p-value < 0.001). We re-analyzed the Affymetrix expression data by applying a RMA summarization and normalization and using the R package limma (67; 68) to fit a linear model containing the following factors: MyoD expression (yes/no), Myog expression (yes/no), and time of differentiation (0, 24, 48, and 96 h). As a positive control set for MyoD or Myog-induced regulation of gene expression we took the top 200 or less genes based on the effect of MyoD or Myog expression, respectively. When needed, accession numbers were converted to Ensembl gene IDs using Idconverter (69).

For the 200 most significantly induced genes, we evaluated whether their promoters contained MyoD or Myog TFBSs according to the ChIP-on-chip data. We expect that the smaller more specific lists would have a higher percent of promoters with true TFBSs (significant on the ChIP-on-chip platform) and therefore likely to contain more significantly over-representated TFBSs in our predictions. We found that as a general trend this is true that the smaller more specific expression lists contain a higher percent of true positives (significant ChIP-on-chip genes) (Additional File 2.1).

2.3.3 Random Data Size Evaluation

We evaluated what would be an appropriate number of random promoters by running a set of 14 experimental promoters against several random set sizes; 500, 1000, 2000, and 4000. For this, the Match cutoff was set to minimize the sum of false positives and negatives. For this test we used a promoter size of 1000 bp before exon 1 and all of exon 1. The larger the random size used the more consistent the number of TFBSs

that were identified (Additional File 2.2), but also the longer the run time. We found a random size of 2000 promoters to be the best trade off between accuracy and speed.

2.3.4 Promoter Size Evaluation

We evaluated an appropriate promoter size for our TFs of interest by taking the *Cao et al.* 2006 expression data top 50 MyoD- or Myog-responsive promoters for the appropriate stimulation (MyoD or Myog) compared to 2000 purely random mouse Ensembl promoters. We varied the promoter size to include exon 1 plus an additional number of bp upstream; 500, 1000, 2000, and 4000. Analysis showed that with a Match setting to minimize false positives a promoter size of 2000 bp + exon 1 was best, whereas with a Match setting to minimize the sum of false positives and negatives a promoter size of 1000 bp + exon 1 was preferable (Additional File 2.3). We continued with a Match setting to minimize the sum of false positives and negatives setting using 1000 bp upstream + exon 1 as our promoter size.

2.3.5 Evaluation of GC Content

To evaluate the effect of GC content we ran purely random Ensembl promoters (the FAST setting of CORE_TF) on all *Cao et al* ChIP data. We then compared that to runs with the option to get random promoters of approximately equal %GC content compared to the experimental set (the Similar %GC option).

2.3.6 Wet-lab Verification of a CORE_TF Predicted Conserved TFBS

To give wet-lab confirmation to the results of the CORE_TF conservation predictions we used the TransFactor kit with double stranded DNA designed on a *LAMA4* (ENSG00000112769) MyoD predicted TFBS conserved between human, chimp, and dog (Figure 2.4). This was an Ensembl genomic alignment run with a Match setting to minimize the sum of false positives and false negatives. The promoter size was defined as 3000 bp upstream of exon 1 and including exon 1. We also included a negative control of the same DNA sequence with four mutations. Recombinant MyoD protein was used to test for binding. For more details on the TransFactor run see the additional material (Additional File 2.4).

2.3.7 CORE_TF Compared to an Existing Program: oPOSSUM

To evaluate our script with existing technology we ran the *Cao et al* 2006 expression data (most significant 20, 50, 100, and 200 genes) through the oPOSSUM website (60). We chose oPOSSUM for comparison since it performs similar analysis and is freely available. We used their custom single site analysis page. Other than setting to mouse, vertebrate JASPAR PWMs, retrieving 1000 bp up and 433 bp downstream (using Ensembl API we calculated this as the average size of exon 1) of the transcription start site, and showing all results, all settings used their defaults. It must be noted that JASPAR only has a PWM for Myf, which represents a TF family including

MyoD and Myog. We also used their number of hits in their background and target genes to run a binomial test in the statistical package R to match our data.

2.3.8 CORE_TF Compared to an Existing Program: ConTra

We also chose to evaluate CORE_TF versus an additional easily viewable cross-species conservation program, ConTra (61). As a test promoter for comparison we used the *LAMA4* (ENSG00000112769) promoter, for which we had a lab verified MyoD TFBS. The ConTra website was run on all default parameters (selecting transcript ENST00000230538), except for looking at 3000 bp upstream instead of 2000 bp upstream (giving a promoter the same size as the CORE_TF run). We looked at the PWM MyoD_Q6_01. This was the only PWM for MyoD available at the ConTra website and the best performing for CORE_TF with this promoter.

2.4 Results and Discussion

2.4.1 CORE_TF Work Flow and Function

We have developed a series of web pages to identify TFBSs in two sequential processes. First, pages 1 to 3 allow a user to predict TFs that regulate a set of co-regulated genes. This is done by identifying TFBSs that are over-represented in the promoters of an experimental (e.g. similar expressed genes from microarray data) compared to a random data set, taking GC content into account if requested. These results are displayed in a sortable table in page 3 (Figure 2.2). Secondly, pages 4 to 6 allow a user to identify specific TFBSs by looking for across species conservation of TFBSs selected from the TFBSs in page 3 and the promoters of page 3. This is done on Ensembl genomic alignments or BLASTz alignments of orthologous promoters provided by Ensembl or the user. Across species conserved TFBSs are displayed in tables (calculations as in Figure 2.3) in page 5 and as aligned promoters in a graphical format (Figure 2.4) in page 6.

Alternatively, if a user did not wish to look at a list of promoters, but just a single promoter they could look purely for cross-species conserved TFBSs by skipping straight to page 4 from page 1. They must then provide which promoter they want to search and a set of TFBSs from a web displayed list. In theory they could paste the sequences conserved in the alignments back into the over-representation pages to find TFBSs over-represented in conserved regions (as opposed to the normal order of looking for conservation with over-represented TFBSs).

2.4.2 Prediction of Over-Represented TFBSs

To evaluate the performance of our tool we first used the *Cao et al* 2006 ChIP-on-chip data as a positive control. We tested whether the promoters in the ChIP pull-down were enriched for the TFBSs for the TFs targeted in the ChIP experiments compared to a random set of promoters. To evaluate the effect of matching promoters for %GC content, CORE_TF was run with a purely random selected set of promoters (FAST option) and a random set of promoters with matched %GC content as controls (similar %GC option). Using both sets of random promoters, CORE_TF found a significant

over-representation (p-value < 0.05, after applying multiple test correction with Benjamini Hochberg in R (70)) for the MyoD PWM MYOD_Q6 in the MyoD bound promoters and the Myog PWM MYOGENIN_Q6 in the Myog bound promoters, in both C2C12 and MDER cells (Additional File 2.5). The MyoD PWM MYOD_Q6_01 was also significant in all MyoD targeted runs except the MDER MyoD with random promoters matched on %GC content.

Strikingly, by ranking TFBSs on p-value, we demonstrate that the target TFs were higher ranked with the %GC matched promoters as control rather than with the purely random set of control promoters (Table 2.1), indicating that improper matching of GC content leads to false positive identification of TFBSs. By evaluating the distribution of p-values for all TFs using both random sets, we observed purely random promoters yield more high and low p-values than a random set of promoters matched on %GC content (Additional File 2.6). Since our target ChIP TFs remained significant when using %GC matched promoters, resulting in a smaller list of significant TFBSs, we believe this method to yield less false positives.

To demonstrate that our algorithm is able to find shared regulatory sites in co-regulated genes identified in expression microarray data we evaluated whether genes for which the expression level increased upon MyoD or Myog activation were enriched for MyoD or Myog TFBSs. We ran sets consisting of the 20, 50, 100, and 200 genes most significantly affected by MyoD or Myog activation versus a random set of approximately equal %GC content (Additional File 2.7). We found significant enrichment of the MyoD_Q6 PWM in all MyoD enriched sets. We also found MYOD_Q6_01 enriched in the top 50 and top 100 MyoD enriched sets. MYOGENIN_Q6 was found enriched in the top 20 Myog enriched set only. Other PWMs for MyoD or Myog and other sets of promoters were not significant or considered "NA" due to 100% of promoters hit in the experimental data. The same data was also run through with the CORE_TF FAST setting. We found that the two settings perform similar, with slightly higher frequencies but slightly less significant p-values when matching on %GC (Figure 2.5). Additionally, as expected the smaller more specific lists generally have higher frequencies and lower p-values than larger, less specific lists (Figure 2.5).

Table 2.1: Cao et al 2006 top ChIP-on-chip predictions with CORE_TF

A. MyoD ChIP-on-chip							
C2C12 MyoD FAST	p-val*	C2C12 MyoD %GC	p-val*	MDER MyoD FAST	p-val*	MDER MyoD %GC	p-val*
AP1_Q6_01	0	MYOGENIN_Q6	1.5E-06	AP1_Q6_01	0	AP4_Q5	0
E2F1DP1_01	0	AP4_Q5	2.3E-06	AP4_Q5	0	AP4_Q6	3.1E-06
E2F4DP2_01	0	E2A_Q2	2.7E-06	COUP_DR1_Q6	0	MYOGENIN_Q6	2.7E-05
E2F_Q4	0	AP4_Q6	8.8E-06	E2F1DP1_01	0	AP4_Q6_01	2.7E-04
E2F_Q6_01	0	**MYOD_Q6**	8.8E-06	E2F4DP2_01	0	AP4_01	1.2E-03
GATA3_01	0	AP4_Q6_01	5.1E-05	E2F_Q4	0	**MYOD_Q6**	4.5E-03
MAF_Q6_01	0	E47_01	1.1E-03	E2F_Q6_01	0	E2A_Q2	5.2E-03
NF1_Q6_01	0	E12_Q6	1.4E-03	MAF_Q6_01	0	LBP1_Q6	2.3E-02
NFE2_01	0	LBP1_Q6	4.0E-03	NF1_Q6_01	0	HEN1_02	7.6E-02
OSF2_Q6	0	E2A_Q6	4.6E-03	NFE2_01	0	TAL1BETAE47_01	7.6E-02
AP4_Q5	7.9E-09	SMAD_Q6_01	2.7E-02	NFKB_Q6	0	**MYOD_Q6_01**	1.7E-01
MYOGENIN_Q6	7.9E-09	**MYOD_Q6_01**	4.4E-02	OSF2_Q6	0	HEB_Q6	2.9E-01
LBP1_Q6	2.2E-08	AP1FJ_Q2	4.5E-02	AP4_Q6	3.5E-09	HELIOSA_01	2.9E-01
AP4_Q6	5.8E-08	AP1_Q4	5.8E-02	GATA3_01	3.5E-09	AP1_Q4	4.0E-01
AP4_Q6_01	6.2E-08	E47_02	9.4E-02	LBP1_Q6	6.5E-08	HNF4_01	4.0E-01
B.Myog ChIP-on-chip							
C2C12 Myog FAST	p-val*	C2C12 Myog %GC	p-val*	MDER Myog FAST	p-val*	MDER Myog %GC	p-val*
AP1_Q6_01	0	AP4_Q5	0	AP1_Q6_01	0	AP4_Q5	0
AP4_Q5	0	AP4_Q6	0	AP4_Q5	0	AP4_Q6	0
AP4_Q6	0	**MYOGENIN_Q6**	0	AP4_Q6	0	**MYOGENIN_Q6**	0
E2F1DP1_01	0	MYOD_Q6	5.0E-06	COUP_DR1_Q6	0	AP4_Q6_01	2.5E-06
MAF_Q6_01	0	AP4_Q6_01	1.1E-05	E2F1DP1_01	0	LBP1_Q6	8.0E-06
MYOGENIN_Q6	0	E2A_Q2	9.0E-04	E2F4DP2_01	0	MYOD_Q6	8.1E-06
NF1_Q6_01	0	AREB6_01	6.9E-03	E2F_Q4	0	E2A_Q2	7.1E-03
NFE2_01	0	MYOD_Q6_01	1.4E-02	E2F_Q6_01	0	MYOD_Q6_01	8.6E-03
OSF2_Q6	0	LBP1_Q6	1.4E-02	LBP1_Q6	0	CLOCKBMAL_Q6	4.5E-02
E2F4DP2_01	4.7E-09	AP4_01	1.9E-02	MAF_Q6_01	0	AP2ALPHA_01	5.7E-02
COUP_DR1_Q6	1.6E-07	AP1_Q4	2.2E-02	**MYOGENIN_Q6**	0	ZEC_01	5.8E-02
AP4_Q6_01	2.3E-07	ZEC_01	2.2E-02	NF1_Q6_01	0	AP2_Q6	7.7E-02
E2F_Q4	1.3E-06	E2A_Q6	7.1E-02	NFE2_01	0	AP4_01	1.1E-01
GATA3_01	6.5E-06	ATF6_01	8.0E-02	OSF2_Q6	0	PPARG_01	1.2E-01
MYOD_Q6	7.5E-06	E47_01	8.2E-02	AP4_Q6_01	6.4E-09	CMYC_02	1.2E-01

CORE_TF predictions on Cao et al 2006 ChIP-on-chip data. Target TFBSs are presented in bold. * = p-values are Benjamini Hochberg corrected. Note: in the MyoD FAST runs MYOD_Q6 and MYOD_Q6_01 had p-values < 0.05 but were not in the top 15 significant TFBSs.

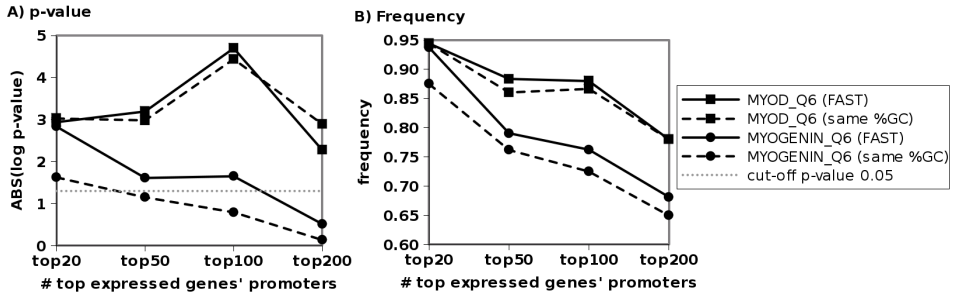


Figure 2.5: Significance of myogenic TFBSs in expression data: The (A) significance (as the absolute value of the log10 p-value) and (B) frequency of MyoD (PWM MyoD_Q6) or Myog (PWM MYOGENIN_Q6) TFBSs in varying number of promoters from genes with increasingly less significant differences in expression upon MyoD or Myog activation are shown. As would be expected, the smaller more significant lists generally have higher frequency and more significant p-values than larger less specific lists.

2.4.3 Orthologous Alignments Versus Genomic Alignments

In many CORE_TF runs we assessed the conserved TFBSs using alignments based on homologous Ensembl promoters as well as Ensembl genomic alignments. Ensembl pairwise alignments can be considered syntenic (they are grouped to make the actual Ensembl synteny blocks) (71). Ensembl orthologs are identified using protein tree calculations (62). The number of promoters aligning and the quality of the alignment to the reference promoter varies tremendously amongst different promoters for both methods (data not shown), but we did not find one method outperforming the other. Synteny does not imply the start of one gene corresponds to the start of a gene in another species. Therefore, this could give poor predictions for TFs that bind and function close to the transcription start site. However, due to many incorrect exon 1 annotations it is also possible that using orthologous promoter alignments may align regions that are not corresponding regions (if an annotation missed exon 1, exon 2 would be annotated as exon 1 and we would instead align to it). Therefore there is not one alignment method that outperforms another to predict conserved TFBSs.

2.4.4 TFBSs Conserved in Orthologous Alignments

The top 10 ranked genes of the Myog-induced genes were inspected for the presence of MYOGENIN_Q6 motifs. To this end, all available orthologs for the mouse genes were retrieved. All conserved TFBSs and their conservation scores are reported in Table 2.2. There are seven promoters which appear to have conserved TFBSs. Four of these promoters (*Chrng*, *Myog*, *Acta1*, and *Tnnc1*) had hits in three or more orthologs. We also inspected the MyoD induced genes presence of MyoD_01 motifs using the same approach and identified two promoters with conserved TFBSs (Table 2.2). Only one promoter was found conserved over three or more orthologs (*Rgs16*). In addition, of the nine across species conserved TFBSs all except *Tnnc1* (not on the array), *Tnnc2*,

Rgs16, and *Nptx1* were found significant in the ChIP-on-chip data. Literature was examined to see if predictions were correct. We found evidence for binding of Myog to *Myog* (72), *Tnni1* (73), and *Chrng* (74). We also found evidence for MyoD binding *Nptx1*, also called *NP1* (75).

Table 2.2: Orthologous conservation of target TFBSs in target genes

A						
Gene Name	GeneID	TF_Name	Tot. Score	Score	#Promos	Length
Chrng	ENSMUSG00000026253	MYOGENIN_Q6	1	1000	5	10
Chrng	ENSMUSG00000026253	MYOGENIN_Q6	1	1000	5	10
Tnnt3	ENSMUSG00000061723	MYOGENIN_Q6	1	1000	2	10
Tnnc2	ENSMUSG00000017300	MYOGENIN_Q6	1	800	2	8
Tnni1	ENSMUSG00000026418	MYOGENIN_Q6	1	800	2	8
Myog	ENSMUSG00000026459	MYOGENIN_Q6	0.83	666.7	5	8
Acta1	ENSMUSG00000031972	MYOGENIN_Q6	0.8	640	4	8
Tnnc1	ENSMUSG00000021909	MYOGENIN_Q6	0.72	720	4	10
Acta1	ENSMUSG00000031972	MYOGENIN_Q6	0.6	480	3	8
B						
Gene Name	GeneID	TF_Name	Tot. Score	Score	#Promos	Length
Rgs16	ENSMUSG00000026475	MYOD_01	1	1200	4	12
Rgs16	ENSMUSG00000026475	MYOD_01	0.5	600	2	12
Nptx1	ENSMUSG00000025582	MYOD_01	0.4	840	2	21
Nptx1	ENSMUSG00000025582	MYOD_01	0.4	480	2	12

Conserved TFBSs for (A) Myog (PWM MYOGENIN_Q6) and (B) MyoD (PWM MYOD_01) from target genes' promoters in expression data. Total score represents a score of conservation from 0 to 1 over the conserved TFBS length. Score represents an additive score over the TFBS. Promos is the number of promoters with the conserved TFBS. Length is the length of the TFBS.

2.4.5 Wet-lab Confirmation of a CORE_TF Predicted Conserved TFBS

To confirm a CORE_TF conserved TFBS in the lab we looked at a MyoD predicted TFBS in the *LAMA4* promoter. Using Ensembl defined genomic alignments we found the matrix MyoD_Q6_01 conserved in human, chimp, and dog (Figure 2.4). Using a recombinant MyoD protein and the TransFactor kit we found significant (p-value 1.5E-35) binding to our target TFBS compared to a mutated one (Additional File 2.4).

2.4.6 CORE_TF Compared to Existing Programs: oPOSSUM

We compared the performance of CORE_TF (using a random set with similar %GC) to oPOSSUM, a webtool with similar objectives as ours. oPOSSUM looks for over-represented JASPAR PWMs in pre-defined species alignments, but is limited to spe-

cific species alignments (e.g. human-mouse) and use of the smaller JASPAR PWM database. We used the previously mentioned expression microarray datasets for the evaluation of both programs performances. Our runs on the oPOSSUM website showed that our binomial test performs similar to their Fisher test (Additional File 2.8). Unlike our frequency observations, the frequency identified by oPOSSUM of TFBS hits in the MyoD induced set did not show the expected high to low pattern (Additional File 2.9). When comparing p-values from the binomial tests for the predictions by the two programs, we see similar patterns between the two programs across the top 20, 50, 100, and 200 genes, but CORE_TF has more significant MyoD predictions and oPOSSUM has more significant Myog predictions (Additional File 2.9). It must be noted that we are only comparing over-represented TFBSs whereas oPOSSUM has already taken conservation into their program at this point which may explain higher sensitivity for Myog promoters. We instead do this on individual promoters and display it graphically in the next step. We believe this graphical representation to be more interpretable.

Since we can do better in one out of two tested TFs without our orthologous promoter conservation we believe CORE_TF to be a superior tool. The two programs differ on several other levels. oPOSSUM only takes Ensembl IDs as input, whereas we also accept nucleotide sequences. We also offer a larger choice of random data sets and conservation methods, as well as the choice to account for GC content. In addition, our number of vertebrate species available is six, all of which can be compared together. oPOSSUM only accepts two species comparisons at a time. For vertebrates oPOSSUM is limited to only human and mouse, both of which are in CORE_TF. In addition, we display our across-species TFBSs in a graphical format, whereas oPOSSUM presents their data in a less intuitive tabular format.

2.4.7 CORE_TF Compared to an Existing Program: ConTra

We also evaluated CORE_TF versus ConTra using the *LAMA4* promoter, for which we had experimental data available, as an example. ConTra is a website to identify and easily view conserved TFBSs in a single cross-species promoter alignment, but cannot look for over-representation in a large promoter set. We found that in CORE_TF genomic alignment predictions there were three MyoD TFBSs conserved between human and chimp and one TFBS conserved between human, chimp, and dog (Figure 2.4). ConTra found the same TFBSs, but also three additional (Additional File 2.10 and data not shown). Two of the three human/chimp CORE_TF conserved TFBSs and the human/chimp/dog CORE_TF conserved TFBS were also found conserved in the macaque in ConTra. CORE_TF did not search for macaque, but it is extremely similar to human and chimp so we believe it would not add much information. However, if a user wanted any Ensembl species added to CORE_TF adding an additional species to the scripts is very simple. It is not surprising the same TFBSs were identified since both programs use Ensembl alignments and TRANSFAC PWMs. ConTra does have the disadvantage of only using human as a reference genome for automated alignment retrievals, whereas CORE_TF can do this for all six species currently installed. Additionally, CORE_TF does not use an Ensembl multi-species defined alignment, but combines many Ensembl pair-wise alignments into one, allowing any number of Ensembl species to be included in one alignment. ConTra does not

display strand specific binding which CORE_TF does by color coding. Additionally, ConTra does not search for over-represented TFBSs in a group of promoters.

2.4.8 Future Efforts

An item that can be improved in the future is our evolutionary scoring algorithm, e.g. by taking into account the confidence of each nucleotide in the PWM. An additional improvement will be to analyze combinations of TFBSs.

2.5 Conclusion

We have developed a tool for identifying over-represented TFBSs in promoters from co-expressed genes aided by the evaluation of cross-species conservation. CORE_TF is easy to use and displays results in tables or graphically allowing for easy interpretation of the results. Our method seems to correctly predict the presence of experimentally verified TFBSs, as shown by our extensive analysis on *Cao et al.* 2006 expression and ChIP-on-chip data and wet-lab confirmation of a MyoD predicted TFBS in the *LAMA4* promoter. We also show improvements over two existing programs (oPOSSUM and ConTra) with greater flexibility in input data, coverage of a larger number of species, more intuitive output, and the option to account for GC content.

Our tool is provided as a web service free to all non-commercial users.

2.6 Availability and Requirements

Project name: CORE_TF

Project home page: *http://www.LGTC.nl/CORE_TF*

Operating system(s): Linux

Programming language: Perl (we used 5.8.4)

Other requirements: TransFac Professional (we used 11.2), BLASTz, sorttable.js, Math::Cephes (Perl module), Apache (we used 1.3.33)

License: GNU General Public License, v3 *http://www.gnu.org/licenses/*

Any restrictions to use by non-academics: none for website use, TransFac Professional license for a local install

2.7 Authors' contributions

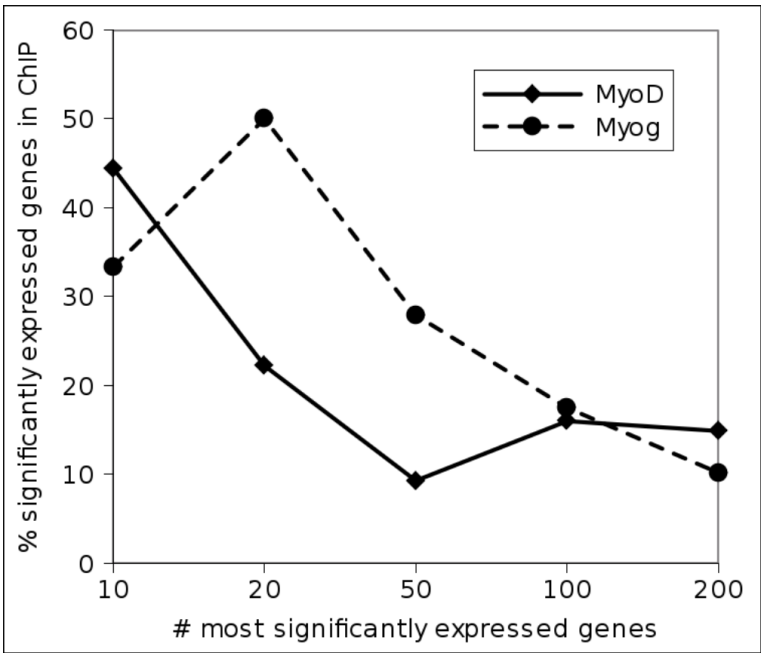
MH, JD, GO, and PH conceived of the primary concepts of the software. MH and MG did the primary programming and debugging. MV performed all primary installations on the web-server and helped in debugging code. MH, MG, and PH performed the software evaluation on expression and ChIP-on-chip data. Wet-lab work was done by MH. Manuscript drafting was done by MH, MG, JD, GO, and PH. All authors read and approved the final manuscript.

2.8 Acknowledgements

We would like to thank Renee de Menezes and Maarten van Iterson for their statistical comments and Ivo Fokkema for his programming and implementation assistance. This work was funded by the Center for Biomedical Genetics (in the Netherlands). PH was supported by a VENI-grant from the Dutch Organization for Scientific Research (NWO grant 2005/03808/ALW).

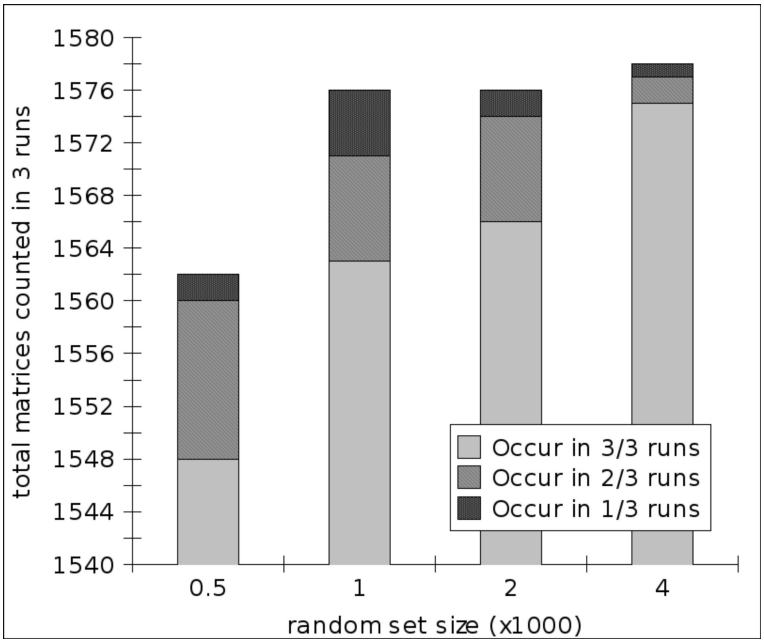
2.9 Additional Files

Additional File 2.1



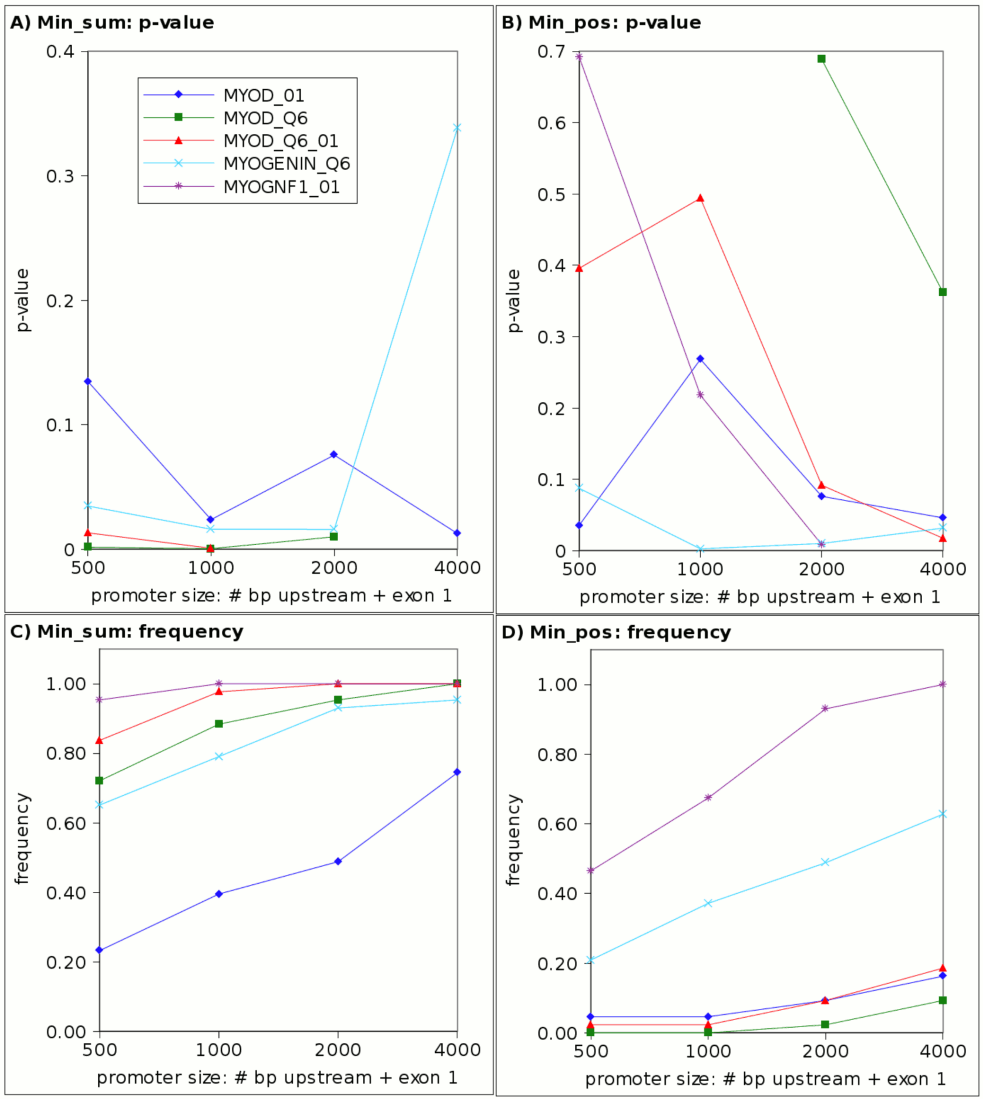
Overlap of most significant expression genes in ChIP-on-chip data. Indicated are the size of the lists for the top expressed genes and the percent of those contained in the significant ChIP-on-chip genes (true-positives). There is a trend that the smaller more selective expression gene lists contain a higher percent of true positives.

Additional File 2.2



Consistency of TF identification in different random set sizes. Indicated are the number of TFs that occur in 1, 2, or 3 out of 3 total runs. As expected, the larger the random set size (500, 1000, 2000, or 4000 promoters) the larger the consistency over runs. However, as indicated by the y-axis scale, this is not a very large effect.

Additional File 2.3



Optimal promoter size. The p-value and frequency of promoters with size 500, 1000, 2000, and 4000 bp and exon 1 with Match settings to minimize false positives (Min_pos) or minimize the sum of false positives and negatives (Min_sum). Overall, we see a promoter of (A) 1000 bp + exon 1 works best for Min_sum runs and (B) 2000 bp + exon 1 works best for Min_pos runs. As expected, (C and D) frequency of TFBSs hit increases as the promoters become larger. For a full color figure see *www.biomedcentral.com/content/supplementary/1471-2105-9-495-s3.tiff*.

Additional File 2.4
TransFactor *LAMA4*-MyoD. Set-up and data analysis of MyoD binding a *LAMA4* promoter derived sequence with the TransFactor kit.

TransFactor confirmation MyoD binds the LAMA4
(ENSG00000112769:ENST00000230538) promoter

Materials:

TransFactor Kit (Clontech product 631956)
Oligos: (ordered from Operon, bring up in TE to 100μm)

LAMA4_MyoD_F	biotin	tgctttcCACCAGCTGTGCgaccttg
LAMA4_MyoD_R		caaggtcgcacagctggtggaaacga
Neg_MyoD_F	biotin	tgctttcC**T**C**G**AG**GA**GTGCgaccttg
Neg_MyoD_R		caaggtcgcac**tc**ct**cga**ggaaagca

* **Nucleotides** are the mutated nucleotides from the original target sequence

Antibodies: Primary: Santa Cruz MyoD (M318): sc760
Secondary: goat anti rabbit IgGHRP from TransFactor Kit
Protein: Recombinant MyoD protein
Plate Reader: BIOTEK Synergy HT

Methods:

Oligo preparation done as:
-mix 10μl forward + 10μl reverse oligo
-place 95°C heat block 10 minutes
-cool on desktop 30 minutes
-mix 20μl with 198μl Mg to make 1μM concentration, vortex briefly
The TransFactor Kit User Manual: V. Colorimetric TransFactor ELISA Procedure is followed with the following additions/changes:
-dilute MyoD antibody 1:100
-dilute goat anti rabbit antibody 1:1000
-step F1: after adding the TMB substrate place directly into the reader
-plate reader protocol:
1. Kinetic 13x5 minute intervals
2. Absorbance
3. Wavelength: 655nm
4. Shake 30s/read

Results:
Measurements over 5 time points:
slope: Tn-T(n-1)

T2-T1 sample	9-26-06		9-29-06		10-6-06	
Neg_MyoD	0.01	0.008	0.003	0.005	0.027	0.028
LAMA4_MyoD	0.021	0.034	0.026	0.03	0.612	0.455
T3-T2 sample						
Neg_MyoD	0.008	0.007	0.001	0.003	0.023	0.022
LAMA4_MyoD	0.019	0.029	0.024	0.024	0.48	0.355
T4-T3 sample						
Neg_MyoD	0.007	0.006	0.001	0.001	0.017	0.02
LAMA4_MyoD	0.017	0.024	0.019	0.02	0.387	0.292
T5-T4 sample						
Neg_MyoD	0.007	0.006	0 0.003	0.017	0.017	
LAMA4_MyoD	0.015	0.02	0.019	0.019	0.322	0.246

Gnumeric spreadsheet Anova single factor results:

Groups	Count	Sum	Average	Variance
measurements	48	3.756	0.07825	0.02247
sample	48	72	1.5	0.25532
day	48	96	2	0.68085

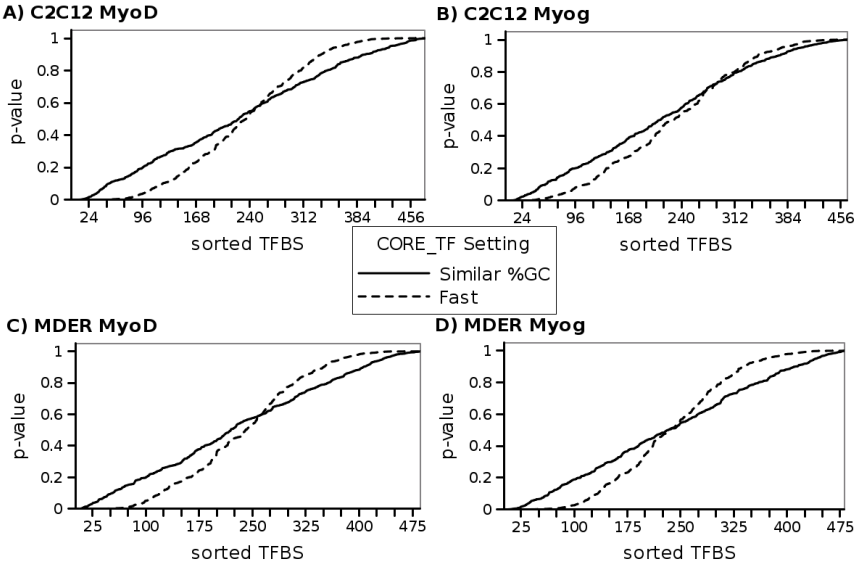
ANOVA

Source of Variation	SS	df	MS	F	P-value	F critical
Between Groups	95.4319	2	47.7160	149.324	1.5E-35	3.06029
Within Groups	45.0561	141	0.31955			
Total	140.488	143				

Conclusion: With a p-value of 1.5E-35 there is a very significant difference in MyoD binging between the negative and target oligos. It is therefore highly likely that the target sequence is a TFBS for MyoD.

Additional File 2.5
Cao_et_al_2006_ChIP_CORE_TF. CORE_TF run results to identify over-represented TFBSs in MyoD/Myog ChIP-on-chip data.
(*http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613159/bin/1471-2105-9-495-S5.xls*)

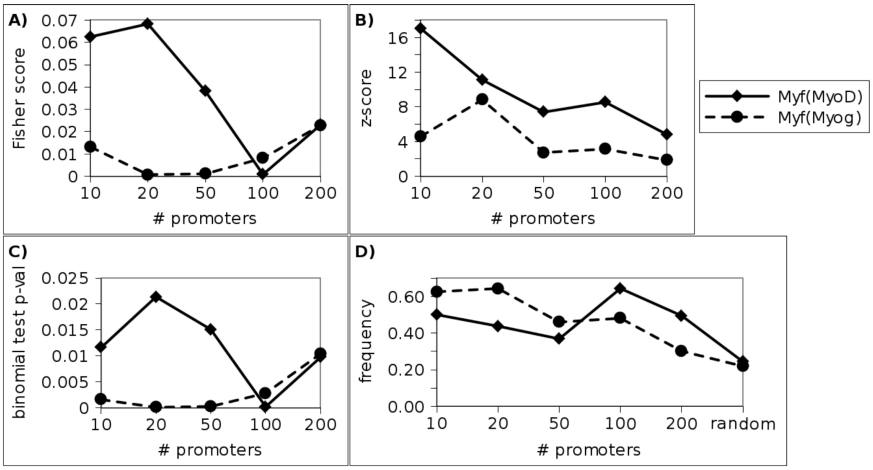
Additional File 2.6



CORE_TF using random FAST runs vs runs with similar %GC. It is visible that in all ChIP-on-chip data tested the runs on purely random Ensembl promoters (FAST runs) has a bias towards high and low p-values while the random set with a similar %GC follows a more normal distribution. This could account for false positives in the FAST runs.

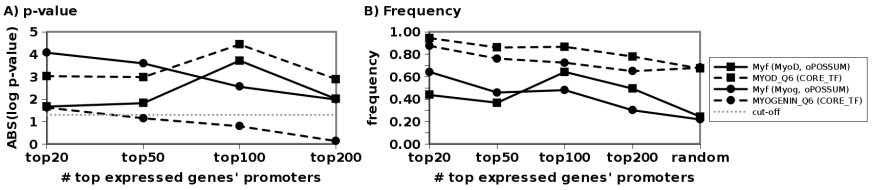
Additional File 2.7
Cao_et_al_2006_expression_CORE_TF. CORE_TF run results to identify over-represented TFBSs in expression array data.
(*http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613159/bin/1471-2105-9-495-S7.xls*)

Additional File 2.8



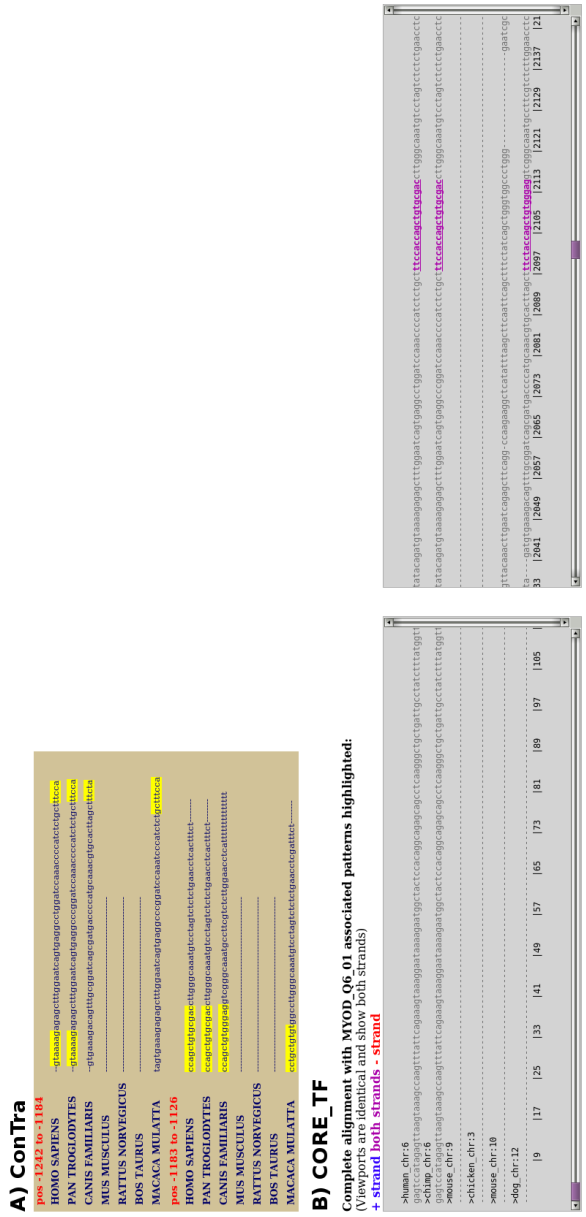
oPOSSUM runs on expression data. Custom oPOSSUM runs using the top 10, 20, 50, 100, and 200 genes from Cao et al 2006 expression data. oPOSSUM supplies (A) Fisher and (B) z-scores. (C) We also used their hits in the experimental and background data to generate a binomial test p-value similar to our program. (D) Frequency of TFBS hits overall declines as we stray from the top hits, as expected, but this is not an entirely smooth curve.

Additional File 2.9



CORE_TF vs oPOSSUM. CORE_TF and oPOSSUM binomial test p-values for the top 20, 50, 100, and 200 genes from Cao et al 2006 expression data for over-expression (A) of MyoD or Myog in the appropriately induced cell line. We see comparable results in the top 20, 50, 100, and 200 sets, but better overall performance in oPOSSUM for Myog and in CORE_TF for MyoD. Frequency (B) of MyoD or Myog hits was also plotted. As expected, the smaller more significant lists generally have higher frequency and more significant p-values than larger less specific lists. Frequency of TFBSs in the promoters was also overall higher in experimental data than random promoters as expected. The oPOSSUM MyoD frequency was the only plot that did not seem concordant.

Additional File 2.10



Identifying MyoD TFBSs conserved in the *LAMA4* promoter with ConTra and CORE_TF. Many conserved TFBSs were found identically between the two programs. Shown here is the most conserved TFBSs found, a MyoD TFBS conserved between human, chimp, and dog in (B) CORE_TF and also macaque in (A) ConTra. Though found by both programs, CORE_TF also identifies the TFBS is on both strands of the DNA. For a full color figure see *www.biomedcentral.com/content/supplementary/1471-2105-9-495-s10.png.*

Chapter 3

Modeling Competition of Transcription Factors for DNA Binding Sites Improves Binding Site Predictions

Matthew S. Hestand[1,2,3], Michael M. Hoffman[1,4+], Ewan Birney[1], Gert-Jan B. van Ommen[2], Johan T. den Dunnen[2,3], Peter A.C. 't Hoen[2]

[1]PANDA group, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.
[2]The Center for Human and Clinical Genetics, Leiden University Medical Center, Postzone S4–0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[3]Leiden Genome Technology Center, Leiden University Medical Center, Postzone S4–0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[4]Graduate School of Life Sciences, University of Cambridge, 17 Mill Lane, Cambridge CB2 1RX, UK.
[+]Current address: Department of Genome Sciences, University of Washington, PO Box 355065, Seattle, WA 98195–5065, USA.

manuscript in preparation

3.1 Abstract

Background: Existing computational methods for prediction of transcription factor binding sites largely ignore competition between transcription factors for binding the same target DNA sequence. We used Sunflower, a program that implements a hidden Markov model to calculate the probability of transcription factor binding to each nucleotide position in a DNA sequence and to account for competition effects. We utilized Sunflower in conjunction with over-representation analysis for predictions of transcription factor binding sites in sets of co-regulated genes.

Results: The validity of our method is demonstrated by the significantly increased probability of binding of transcription factors targeted in chromatin immunoprecipitation (ChIP) experiments in the immunoprecipitated DNA sequences. This was established for different transcription factors (MyoD, Myog, p53, and STAT1) and technological platforms (ChIP-chip, ChIP-paired end ditag sequencing, and ChIP-seq). We observed that the *a priori* binding probabilities were dependent on the DNA sequence characteristics. It is therefore essential to match the background DNA sequence to the sequence regions of interest, *e.g.* separate CpG islands and CpG deserts.

Conclusion: With this method, it is possible to predict transcription factor binding sites in sets of co-regulated genes and predict transcription factors that co-regulate gene expression with transcription factors targeted in ChIP experiments. Our method outperforms other approaches that do not account for competition between transcription factors. Furthermore, our approach models the true biological state more realistically in which transcription factors may compete for similar genomic regions.

3.2 Introduction

Several laboratory methods for the identification of transcription factor (TF) binding sites (TFBSs) are available. These include luciferase reporter assays and chromatin-immunoprecipitation coupled with either a microarray assay (ChIP-chip) (34), paired-end ditag sequencing (ChIP-PET) (11), or next-generation sequencing (ChIP-seq) (35). However, all of these methods are time-consuming and expensive. In addition, they tend to identify binding regions rather than binding sites at single base pair resolution. To identify TFBSs more quickly, less expensively, and more precisely, several computational tools have been developed. However, also computational identification of TFBSs can be a cumbersome process. A TFBS may be less than 12–14 base pairs (bp) long and have a fairly loose consensus sequence (49). Position weight matrices (PWMs), which summarize experimental information on the sequence preference of TFs, are commonly used in the search for TFBSs of known TFs (50). Two leading databases of PWMs are TRANSFAC (51; 52) and JASPAR (53; 54). TRANSFAC is larger with 834 PWMs in total (release 11.4, December 2007), compared to 123 in JASPAR. However, one may use the complete database of JASPAR PWMs for free, while licensing fees are required to use the complete database of TRANSFAC Professional PWMs. Existing programs, such as Match (55; 51), identify TFBSs by evaluating the nucleotide similarity of the PWM with the genomic sequence of interest. A TFBS is predicted when the similarity score passes a threshold.

To increase prediction accuracy, reported PWM alignments are usually further filtered. Several methods take information on the evolutionary conservation of TFBSs into account (76; 61; 77; 60). Another commonly used approach is to search for shared TFBSs in co-regulated genes as it is presumed that similarly expressed genes have common regulators (57). A binomial or analogous statistical test is frequently used to test whether the number of TFBSs predicted in the sequences of interest is statistically higher (*i.e.* erniched) than in a random group of genomic sequences (76). These methods are implemented in several web applications, such as ConTra (61) and COTRASIF (77) for conservation, PSCAN (78), Asap (79), and OTFBS (80) for over-representation analysis, and CORE_TF (76) and oPOSSUM (60) for both.

Few PWM-based algorithms model competition between different TFs. Models have been used to identify TFBSs in insects (81; 82), such as *Drosophila*, which take into account competition. Segal *et al.* 2008 (81) use a competitive model that also requires TF expression levels, but with the aim to predict target gene expression levels. Sinha 2006 (82) take into account competition, but aim at predicting unknown binding motifs. To our knowledge, the first model to address TF binding competition in vertebrates for known TFs is Sunflower (56). Sunflower uses a hidden Markov model which assumes steric hindrance between TFs for the same DNA sequence. This is accomplished by permitting a single path through the model to traverse only one PWM at a time, disallowing TFs to bind to the same place at the same time. Sunflower sums the probabilities of all possible paths through the model using the forward-backward algorithm (83), resulting in a posterior probability per nucleotide position. This probability thus accounts not just for the PWM of that TF, but also for competitive effects due to overlap with PWMs of other TFs. The multitude of different circular paths (from unbound via bound back to the unbound state) gives Sunflower its name, as each path can be represented by one flower leaf attached to

the heart of the Sunflower. For simplicity we declare the DNA region to be bound by the TF when the probability exceeds a given threshold. In the current paper, we use Sunflower in conjunction with enrichment analysis to identify TFBSs.

When searching for enriched TFBSs, it is essential to select an appropriate group of background sequences. It is common practice to select a group of random promoter sequences. However, different classes of promoters may exist. Some promoters, such as those containing a TATA box, have transcription start sites with very defined specific locations, whereas others may contain broad transcription start sites with multiple peaks of transcription (8). The latter promoters are more likely to contain a CpG island (8; 9). Different classes may have different binding affinities for TFs. Furthermore, if the abundance of T/A or G/C nucleotides in target experimental sequences is different from the background sequence set, the estimation of the binding probabilities and frequencies may be incorrect, in particular for PWMs with skewed nucleotide contents. An example of this is finding the GC-enriched PWM for Sp1 over-represented when not selecting appropriate GC content background sequences (84; 85). CpG islands have a higher GC content by definition (10) so one means of separating sequences on their GC content is by discriminating on CpG content as CpG islands or CpG deserts. Incorporating GC and/or CpG content into predictions for both *de novo* motifs and known PWMs has previously been shown to improve results (86; 87; 76; 85). Since CpG islands and CpG deserts have potentially different promoter binding behavior, we followed the philosophy of Roider *et al.* (85) and considered CpG desert and CpG island sequences separately, as well as demonstrate their different behaviors in our model.

In this paper, we evaluated and optimized Sunflower in conjunction with our enrichment algorithms using ChIP-chip, ChIP-PET, and ChIP-seq data where we knew in advance which TF should bind the target sequences. In addition to rediscovering the targeted TFs, we discovered potential co-regulators with over-represented TFBSs in the regions bound by the targeted TF.

3.3 Results

3.3.1 Optimal PWMs

We set out to improve computational TFBS predictions by accounting for competition between TFs, which is not done in existing PWM-based methods (55; 51; 60; 76). The program Sunflower (56) was used to determine the binding probability of a TF at a specific nucleotide position. We considered a TF bound to a specific nucleotide if the probability at that nucleotide position exceeded 0.1. Subsequently, a binomial test was used to evaluate the enrichment of TFBSs in a selected set of genomic regions over a set of background sequences. To demonstrate the validity of our approach, we evaluated whether genomic regions bound by TFs, as determined in ChIP experiments, were significantly enriched for TFBSs for the TFs targeted in the ChIP procedure. Initial ChIP identified regions were from MyoD and Myog immunoprecipitation experiments (66).

Since (partially) redundant PWMs representing the same TF may compete for each other, it is essential to choose an optimal selection of non-redundant PWMs. Given the competition element, our method can also be used to select the best PWM

for a given TF. We took PWMs, including the MyoD and Myog PWMs available, from both TRANSFAC and JASPAR (MyoD/Myog are represented by the Myf PWM in JASPAR, Additional File 3.1). Using the MyoD or Myog bound promoter sequences (±1000 bp around the transcription start site), identified by ChIP-chip, and random promoter sequences retrieved and sorted by CpG content, we ran Sunflower to calculate the binding probability of each nucleotide in a sequence by each PWM and evaluated the enrichment of each PWM.

We saw in both CpG deserts and CpG islands that MyoD/Myog TRANSFAC PWMs had lower p-values and thus higher ranks than the JASPAR Myf PWM (Table 3.1). It should be cautioned that we only compared TRANSFAC and JASPAR MyoD/Myog PWMs for these specific MyoD/Myog ChIP-chip experiments. TRANSFAC PWMs may not have greater significance than JASPAR PWMs for other TFs and experimental sequences. None of the TRANSFAC PWMs were consistently better than the others. To better identify which TRANSFAC MyoD/Myog PWMs were optimal we took a slightly larger selection of PWMs (Additional File 3.1) and observed their performances at a range of TF concentrations in the Sunflower model. This was done by adjusting the prior probability of the unbound state, which is inversely correlated to the concentration of TFs in the model. We adjusted the prior probability of the unbound state from the default of 0.9, to also 0.95, 0.985, 0.99, and 0.999. When increasing the prior probability of the unbound state, one MyoD PWM (V$MyoD_Q6_01) and one Myog PWM (V$MYOGNF1_01) remained significant, while the significance of other PWMs for MyoD or Myog decreased (Figure 3.1). These same two PWMs are also used in TRANSFAC's non-redundant PWM set. Since a higher prior probability of the unbound state invokes stronger competition we believe these two PWMs drive the other MyoD/Myog PWMs out by competition. This also indicates that increasing the stringency by raising the prior probability of 0.999 is best-suited for identification of true TFBSs. We therefore performed the remainder of the analysis with the V$MyoD_Q6_01 and V$MYOGNF1_01 PWMs and a prior probability of the unbound state fixed to 0.999. Using TRANSFAC's non-redundant PWM set, we extended the size of our PWM selection to a total of 102 PWMs (Additional File 3.1). With a larger selection of PWMs, each nucleotide has more TFs competing to bind it and we therefore reduced the threshold to declare a bound nucleotide in the binomial tests from 0.1 to 0.01 for the remaining analyses.

3.3.2 Background Set Analysis

When looking for enrichment of a factor in an experimental compared to a background set, the use of different background sets can have a major impact on reported significance (88; 89). We studied different ways to select experimental sequence regions and random sequences and evaluated the effect on the prediction of MyoD and Myog binding sites in promoters of genes identified in the MyoD/Myog ChIP-chip experiments. For experimental sequences we used all Ensembl (3) promoters, or only promoters from presumed better annotated Ensembl genes (minimum 5′ UTR of 40 bp). As background sequences, we tested random sets of Ensembl promoters, random better annotated Ensembl promoters, promoters that were negative in the ChIP-chip experiments, and random genomic sequences (Figure 3.2). Greatest significance was found when random genomic regions were used as background, confirming the *a priori*

Table 3.1: JASPAR versus TRANSFAC PWMs

ChIP-chip target	PWM	CpG deserts p-value	rank	CpG islands p-value	rank
MyoD	V$MYOD_01	1.84×10^{-4}	7	6.47×10^{-1}	31
MyoD	V$MYOD_Q6	7.16×10^{-5}	5	1.77×10^{-6}	1
MyoD	V$MYOD_Q6_01	0	1*	4.64×10^{-2}	11
MyoD	Myf	1.14×10^{-3}	11	8.77×10^{-2}	14
Myog	V$MYOGENIN_Q6	2.62×10^{-4}	7	4.33×10^{-4}	2
Myog	V$MYOGNF1_01	5.76×10^{-5}	6	6.45×10^{-1}	37
Myog	Myf	1.41×10^{-1}	15	3.62×10^{-3}	5

*tied for 1st place with 2 other PWMs.

JASPAR (Myf) and TRANSFAC (V$MYOD_01, V$MYOD_Q6, V$MYOD_Q6_01, V$MYOGENIN_Q6, V$MYOGNF1_01) PWMs p-values and ranks (within p-values, sorted on significance) vs all other 49 PWMs and the unbound state for binomial tests on CpG-separated MyoD/Myog ChIP-chip data.

assumption that TFs are more likely to bind near genes. There were no large differences between promoters from normal Ensembl genes and more confidently annotated Ensembl genes nor between randomly selected promoters and promoters shown to be negative in the ChIP-chip experiments. We continued with what we considered to be theoretically best, using random Ensembl promoters for genes identified in promoter microarray experiments and random genomic regions for ChIP-PET or ChIP-seq experiments.

3.3.3 Testing for Enrichment With Four ChIP Data Sets

To test the overall validity of our method, we used the larger set of 102 PWMs with four data sets: the previous MyoD and Myog ChIP-chip sequences, sequences from ChIP-PET with p53 (11), and STAT1 ChIP-seq sequences (35). For ChIP-chip analysis we used random promoters (separated on CpG content) as background sequences and for ChIP-PET and ChIP-seq we used random genomic sequences (separated on CpG content) as background sequences.

We first evaluated the significance of the MyoD and Myog PWM in the MyoD/Myog ChIP-chip sequences, and compared their p-values with those of other PWMs. The MyoD PWM was highly significant in the MyoD-immunoprecipitated promoters classified as CpG deserts and also significant (p-value = 0.01) in CpG islands (Table 3.2). Results for the Myog PWM in Myog-immunoprecipitated promoters were overall less significant, but were still significant (p-value <0.05) for CpG deserts (Table 3.3). The TFBSs most significantly enriched in the MyoD/Myog ChIP regions (Tables 3.2 and 3.3), could be binding sites for co-regulators, binding to the same genomic regions as the targeted TFs. We evaluated whether there was literature evidence for this, using the text mining tool Anni (90). Anni uses concept profiles, which summarize the literature context in which a term, such as a biological process or gene, is mentioned in. Out of 10,850 potential Gene Ontology (GO) (91; 92) biological processes terms, the top 10 TFs in MyoD CpG deserts, MyoD CpG islands, and Myog CpG deserts were most strongly associated with the concept profile for myogenesis, the process

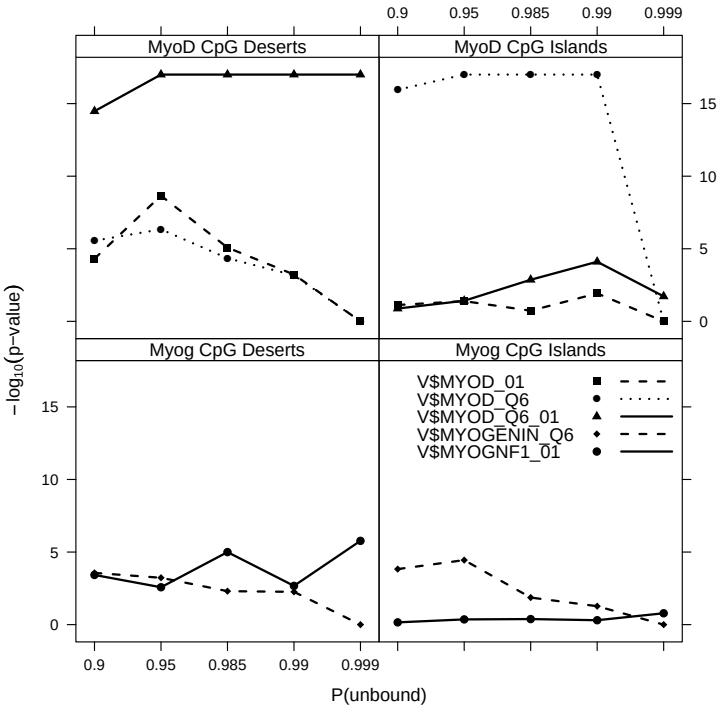


Figure 3.1: Varying the prior probability of the unbound state: Plots of the $-\log_{10}$(p-value from binomial tests) from MyoD PWMs and Myog PWMs in MyoD/Myog ChIP-chip data, sorted on CpG content, when varying the prior probability of the unbound state, P(unbound), in the Sunflower model. Displayed p-values have a minimum of 1×10^{-17}.

in which MyoD and Myog are involved. In Myog CpG islands, myogenesis had the sixth best match. Since MyoD itself contributed considerably to the association with myogenesis, we excluded MyoD and performed the same analysis. Still, myogenesis ranked at position 21 or better for all biological processes. For many TFs predicted as potential co-regulators in myogenesis Anni found co-occurrences of TFs and "myogenesis" in the same MEDLINE abstracts, suggesting potential involvement of these TFs in myogenesis (Tables 3.2 and 3.3).

In the p53 ChIP-PET sequences, we found p53 TFBSs to be highly enriched in ChIP regions classified as CpG deserts and CpG islands (Table 3.4). Similarly, through a literature search to identify potential co-regulators of p53, we found cell cycle arrest and tumor suppressor activity among the best matching biological processes. These are also well-established functions for p53 itself.

We also found STAT1 TFBSs to be highly significant in both ChIP regions sorted as CpG deserts and CpG islands (Table 3.5). In summary, we have validated our approach through prediction of experimentally determined TFBSs and predicted many potential TFBSs for co-regulators in four independent datasets with widely different

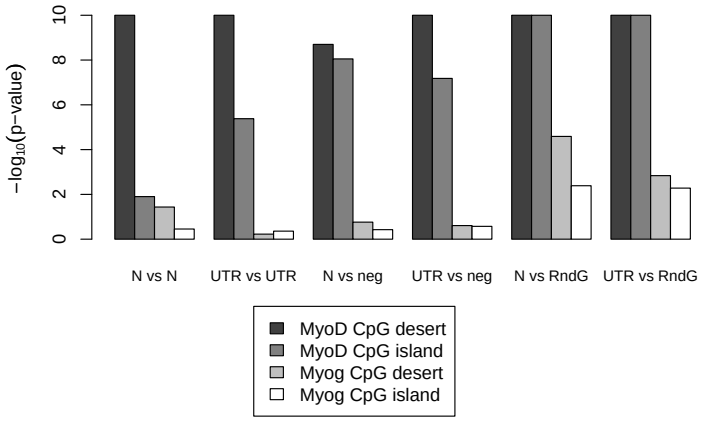


Figure 3.2: Evaluation of background sequences: Plots of the $-\log_{10}$(p-value from binomial tests) from CpG stratified MyoD and Myog ChIP-chip data using several experimental and background sequence types: promoters from normal Ensembl genes (N), promoters from Ensembl genes with at least 40 bp of 5′ UTR (UTR), negative ChIP-chip promoters (neg), and random genomic sequence (RndG). Displayed p-values have a minimum of 1×10^{-10}.

characteristics.

3.3.4 Excluding PWM Nucleotide Composition Bias

To exclude the possibility that the significance of PWMs was caused by nucleotide composition rather than the actual sequence, we replaced the MyoD PWM in the larger selection of 102 non-redundant PWMs with either a PWM in which the last half of the PWM was moved to the first half, or with a PWM in which the nucleotides were placed in completely random order. As expected, the unaltered MyoD PWM was much more significant than the shuffled MyoD PWMs (Table 3.6). We also shuffled the p53 PWM similarly and found the unshuffled p53 PWM to be more significant than the shuffled p53 PWMs (Table 3.6).

3.3.5 Comparison to Existing Programs

For a comparison to other methods, we took the MyoD/Myog ChIP-chip promoters as input and applied them to two other programs that also use statistical tests for enrichment of TFBSs: CORE_TF (76) and PSCAN (78). We tried to keep the same settings for all programs as close as possible. We compared the p-values provided by each method and found that our Sunflower based method had the most significant p-value in three out of the four sets of ChIP sequences (Figure 3.3). CORE_TF and our Sunflower method use the same binomial test for over-representation, but

Table 3.2: Top 10 significant PWMs from MyoD ChIP-chip sequences and literature evidence

PWM	CpG deserts p-value	literature evidence*
V$AP2ALPHA_01	0	Y
V$KROX_Q6	0	Y
V$MYOD_Q6_01	**0**	**Y**
V$SP1_Q2_01	0	Y
V$ZNF219_01	0	-
V$E2A_Q2	1.00×10^{-10}	Y
V$SREBP_Q6	1.00×10^{-10}	N
V$HIC1_02	2.09×10^{-8}	N
V$PAX5_01	4.22×10^{-8}	Y
V$HEN1_01	1.47×10^{-7}	Y
PWM	CpG islands p-value	literature evidence*
unbound	0	
V$IRF_Q6	4.46×10^{-6}	Y
V$E2A_Q2	2.11×10^{-4}	Y
V$KAISO_01	3.71×10^{-4}	N
V$TAL1BETAE47_01	9.14×10^{-3}	Y
V$SRY_02	1.09×10^{-2}	Y
V$NF1_Q6_01	1.20×10^{-2}	N
V$MYOD_Q6_01	1.26×10^{-2}	**Y**
V$TBX5_01	1.36×10^{-2}	Y
V$AP1_Q4_01	2.52×10^{-2}	Y

Significance is determined by p-values from a binomial test. *Literature evidence is based on co-occurrence of the TF's concept profile with the concept profile for "myogenesis." A "-" indicates a defined literature-based concept that did not have a profile.

PSCAN uses a z-test. To compare between these we sorted p-values in descending significance and reported the rank of the target PWM compared to the total number of PWMs. We found that our Sunflower based method performed consistenly best on the evaluated datasets (Figure 3.3), with most pronounced performance on the MyoG CpG islands.

3.4 Discussion

We present a new method for the prediction of enriched TFBSs. The method includes the use of Sunflower, a program that calculates the probabilities of TF binding to nucleotide sequences with a hidden Markov model. These probabilities reflect kinetic rates for TF binding. Most current applications do not take competition between TFs into account, while our system naturally incorporates the competition of different TFs for the same site. In theory, this method should better model the real biological

Table 3.3: Top 10 significant PWMs from Myog ChIP-chip sequences and literature evidence

PWM	CpG deserts p-value	literature evidence*
V$LHX3_01	0	N
V$MYOD_Q6_01	0	Y
V$SP1_Q2_01	0	Y
V$ZNF219_01	0	-
V$KROX_Q6	1.00×10^{-10}	Y
V$AP2ALPHA_01	8.00×10^{-10}	Y
V$E2A_Q2	1.80×10^{-9}	Y
V$HIC1_02	2.70×10^{-9}	N
V$SREBP_Q6	3.70×10^{-9}	N
V$HEN1_01	7.13×10^{-7}	Y
PWM	CpG islands p-value	literature evidence*
unbound	0	
V$AP4_01	2.64×10^{-3}	N
V$HAND1E47_01	4.39×10^{-3}	Y
V$IRF_Q6	9.42×10^{-3}	Y
V$E2A_Q2	1.87×10^{-2}	Y
V$TAL1BETAE47_01	1.92×10^{-2}	Y
V$PAX6_Q2	2.79×10^{-2}	Y
V$SOX9_B1	2.80×10^{-2}	Y
V$TBX5_01	4.57×10^{-2}	Y
V$EVI1_03	4.61×10^{-2}	N

Significance is determined by p-values from a binomial test. *Literature evidence is based on co-occurrence of the TF's concept profile with the concept profile for "myogenesis." A "-" indicates a defined literature-based concept that did not have a profile. Myog ChIP-chip data had p-values of 3.67×10^{-2} (30th most significant PWM) and 3.52×10^{-1} (36th most significant PWM) for V$MYOGNF1_01 in CpG deserts and CpG islands respectively.

system. We demonstrated its validity and improvement over existing methods by correct prediction of TFBSs in ChIP regions. Besides the rediscovery of targeted TFs by our method, the method can identify potential co-regulators in the ChIP regions, as evidenced by literature searches. This could be especially useful to identify DNA binding regulatory elements in ChIP-based methods with an antibody directed against proteins that do not bind the DNA directly (*e.g.* CBP or p300 (93)). The system could also be used to derive TFs common to a set of co-regulated genes identified in expression data.

Besides identifying TFBSs from PWMs, this method can be used to evaluate PWMs themselves. By shuffling the order of the nucleotides (Table 3.6) we could evaluate how much of a PWMs performance is based on the sequence compared to mere nucleotide content. We see shuffling PWMs results in less significant p-values but not complete insignificance, indicating that the nucleotide content alone

Table 3.4: Top 10 significant p53 ChIP-PET PWMs

CpG deserts		CpG islands	
PWM	p-value	PWM	p-value
unbound	0	unbound	0
V$HEN1_01	0	**V$P53_02**	**0**
V$NFY_Q6_01	0	V$BACH2_01	8.21×10^{-3}
V$P53_02	**0**	V$POU6F1_01	1.05×10^{-2}
V$SRF_Q6	3.00×10^{-9}	V$TEL2_Q6	1.26×10^{-2}
V$NRSF_Q4	5.20×10^{-7}	V$AP1_Q4_01	1.53×10^{-2}
V$AP1_Q4_01	3.69×10^{-5}	$EGR1_01	1.64×10^{-2}
V$PPARA_02	4.64×10^{-5}	V$AP2ALPHA_01	2.10×10^{-2}
V$VDR_Q3	1.95×10^{-4}	V$USF_Q6_01	2.45×10^{-2}
V$STAF_02	3.07×10^{-4}	V$ZEC_01	3.00×10^{-2}

Significance is determined by p-values from a binomial test.

Table 3.5: Top 10 significant STAT1 ChIP-seq PWMs

CpG deserts		CpG islands	
PWM	p-value	PWM	p-value
V$AP1_Q4_01	0	unbound	0
V$BACH2_01	0	V$SP1_Q2_01	0
V$STAT1_01	**0**	**V$STAT1_01**	1.06×10^{-8}
V$SP1_Q2_01	8.00×10^{-10}	V$KROX_Q6	8.73×10^{-7}
V$ZNF219_01	4.50×10^{-9}	V$ZNF219_01	7.35×10^{-5}
V$PPARA_02	1.54×10^{-8}	V$STAF_02	1.37×10^{-3}
V$HEN1_01	1.60×10^{-7}	V$P53_02	2.47×10^{-3}
V$NRSF_Q4	2.33×10^{-7}	V$CREB_Q4_01	3.06×10^{-3}
V$USF_Q6_01	8.99×10^{-6}	V$HLF_01	4.53×10^{-3}
V$HIC1_02	1.44×10^{-3}	V$ZEC_01	4.53×10^{-3}

Significance is determined by p-values from a binomial test.

Table 3.6: Shuffling the MyoD and p53 PWMs

TF & Sequence CpG content	Normal PWM	Flip first/last half PWM	Randomized PWM
MyoD CpG desert	0 (1*)	2.78×10^{-6} (12)	4.4×10^{-5} (13)
MyoD CpG island	1.26×10^{-2} (8)	1.00 (88)	1.00 (91)
p53 CpG desert	0 (1**)	0 (1**)	1.89×10^{-2} (18)
p53 CpG island	0 (1***)	3.32×10^{-2} (10)	4.04×10^{-1} (52)

Indicated for each condition: p-value and rank (within parenthesis) out of 102 PWMs and the unbound state (sorted on decreasing significance). *tied for 1st place with 4 additional PWMs. **tied for 1st place with 3 additional PWMs. ***tied for 1st place with 1 additional PWM.

has some predictive value. Besides evaluating a PWM itself, we can compare PWMs for the same TF to find the best performing PWM, as hinted by our MyoD and Myog comparisons (Figure 3.1). The best performing PWMs for MyoD and Myog had the

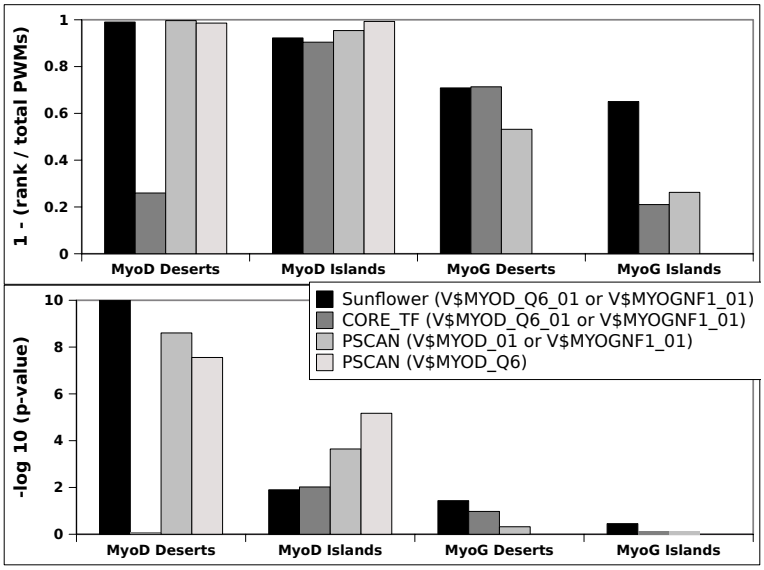


Figure 3.3: We compared our Sunflower results to the programs CORE_TF and PSCAN. The top chart plots rankings of target PWMs in over-representation results between different programs, as indicated by 1 - (rank / total number of PWMs). Therefore, the closer to one the better the performance. The lower chart plots the $-\log_{10}$(p-value) from each method.

longest consensus sequence. This may reflect a bias in the Sunflower algorithm. The longer a PWM is the longer the path in the Sunflower model will be, resulting in a larger posterior probability of TF binding. This is an issue that should be addressed in a future version of Sunflower.

The issue of using appropriate background sequences when searching for over-represented TFBSs has been addressed before (85), but many programs, such as PSCAN, still do not take into account background set differences. We found that for MyoD our predictions performed relatively similar to PSCAN that did not take into account background CpG content, but for Myog our method did perform better (Figure 3.3). To see how well our system would work if we did not separate on CpG content we analyzed the MyoD/Myog data without sorting experimental or background sequences on CpG content. Without sorting we still find the MyoD PMW among the top ranked PWMs (near 0 p-value and tied for first place in ranking). However, we found poorer performance for Myog with a rank of 45 out of 103 (including the unbound state, p-value of 2.08×10^{-1}). In line with the suggestion by Roider *et al.* 2009 (85), we recommend that CpG islands and CpG deserts be treated separately.

Another argument for separating CpG deserts and CpG islands is their inherently different binding properties. An indiciation for this is that, after shuffling the nucleotide sequence in the PWMs, more significance was retained in CpG deserts than CpG islands. This is probably not due to a GC bias in PWMs since most PWMs had near 50% GC content (Additional File 3.2). In the evaluation of the MyoD,

Myog, p53, and STAT1 datasets, we generally observed higher significane levels in CpG deserts than CpG islands. Finding higher significance for TF binding in CpG poor compared to CpG rich regions has been reported previously (85). In line with this, the unbound state is highly significant in CpG islands but not in CpG deserts (with the exception of p53). This reflects a higher overall likelihood of TF binding in CpG deserts than CpG islands.

Different from the effect of matching on CpG content, we found no large influences of the promoter type. However, the TF binding properties of promoter regions were clearly distinct from those of genomic regions (Figure 3.2). We therefore recommend that the background sequences should be matched as closely as possible to the experimental sequences: random promoters for genes identified in expression profiling or promoter microarray experiments and random genomic regions for ChIP-PET or ChIP-seq experiments. This approach parallels the over-representation approaches suggested for matching appropriate background sets to differential gene expression sets from microarray experiments to identify enriched GO terms (88; 89).

To conclude, we have developed a new method for the prediction of enriched TFBSs. Its validity was confirmed by the rediscovery of TFBSs for different TFs (MyoD, Myog, p53, and STAT1) targeted in ChIP-chip, ChIP-PET, and ChIP-seq experiments. In a novel step, we have accounted for TF binding competition in our method with the Sunflower algorithm. Sunflower has the potential to model TF binding in a much more realistic way than its predecessors and should be used, in conjunction with this work, more extensively in the future. By using a model more representative of the actual biological state, where TFs compete for binding in the same regions of DNA, this method will prove valuable in analyses of a variety of ChIP and expression applications. The selection of proper background set also remains an important issue for methods that investigate the enrichment of TFBSs.

3.5 Materials and Methods

3.5.1 Obtaining Experimental ChIP Sequences

We used four data sets for evaluation purposes: MyoD and Myog ChIP-chip data (66), p53 ChIP-PET data (11), and STAT1 ChIP-seq data (35). MyoD/Myog ChIP-chip results were from a differentiating mouse myoblast cell line (C2C12). Positive ChIP-chip lists were identified as those promoters enriched for MyoD or Myog, as defined by Cao *et al.*, 2006 with a p-value below 0.001. The third data set was a p53 chromatin-immunoprecipitated human colorectal cancer cell sample coupled with a PET sequencing approach. We used the 542 genomic loci identified and reported for p53 in Wei *et al.*, 2006's supplemental table 4 for p53 positive ChIP-PET regions. The fourth data set was a STAT1 chromatin-immunoprecipitated IFN-γ-stimulated human HeLa S3 cells coupled with next-generation sequencing on an Illumina 1G system. We used the ChIP-seq regions from their Supplemental Data 1 (file "STAT1_hg18_IFNg_ht11.peaks.txt") that contained over 500 sequencing reads as positive ChIP-seq regions.

For the MyoD/Myog ChIP-chip data, we converted to Ensembl (3) stable identifiers with Idconverter (69) and used the Ensembl Perl API to retrieve promoter sequences, defined as 1000 bp before and 1000 bp after each transcription start site.

As experimental sequences we also tested promoters from Ensembl genes with highly confident annotation (with at least 40 bp of 5′ UTR).

The p53 ChIP-PET and STAT1 ChIP-seq data were of variable lengths, but due to constraints in Sunflower's reporting and testing process we needed sequences of identical length. Therefore we truncated or expanded regions on both ends to a size of 1186 bp (the mean original PET sequence size) and retrieved sequences with the Perl API scripts. ChIP-seq regions were on average 2626 bp, but we chose to use the same length as the ChIP-PET data to give a more precise binding target and to make background sets comparable.

Unless otherwise stated, all ChIP regions were separated on CpG content (classified as CpG deserts or CpG islands) using the EMBOSS suite (newcpgreport) (94).

3.5.2 Obtaining Random Data Sequences

For random sequences, we tried promoters from random Ensembl genes, random Ensembl genes with at least 40 bp of 5′ UTR, negative ChIP-chip genes (genes within the worst 1500 p-values), and random genomic regions. All sequences were retrieved with the Ensembl Perl API and (unless otherwise stated) separated by CpG content using the EMBOSS suite (newcpgreport). For random genomic regions we used the same sequence length as the corresponding experimental sequences. After CpG sorting, we arrived at random sets consisting of 200 regions.

3.5.3 JASPAR and TRANSFAC PWM Selections

For a Sunflower usable format we converted a TRANSFAC .dat file to a JASPAR-style format with a custom Perl script.

JASPAR has solely the Myf family (representative for both MyoD and Myog) PWM, but TRANSFAC Professional has 3 separate MyoD and 2 separate Myog PWMs. We made a selection of PWMs with a mix of TRANSFAC and JASPAR PWMs (including all Myf/MyoD/Myog PWMs). We also made a selection of TRANSFAC only PWMs (including all MyoD/Myog PWMs), and a larger more inclusive selection of 102 non-redundant TRANSFAC-only PWMs (from the TRANSFAC vertebrate non-redundant minimum false positives PWM set, including only one MyoD PWM, V$MYOD_Q6_01, and one Myog PWM, V$MYOGNF1_01). Full lists of all PWM selections are in Additional File 3.1.

For each selection of PWMs, the unbound state was trained with emission probabilities from the background distribution of unambiguous nucleotides in the sequenced genome of the target species (human NCBI36 or mouse NCBIM37 reference genomes) using the fastacomposition script of the exonerate package (95) and Sunrecompose (from Sunflower).

3.5.4 Setting the Prior Probability of the Unbound State

We adjusted the Sunflower model parameter of the prior probability of the unbound state, which essentially represents the concentration of TFs in the model. Initially the selection of mixed TRANSFAC/JASPAR PWMs used a prior probability of the unbound state fixed to the default 0.9. To analyze varying concentrations of TFs in the model, we varied the prior probability of the unbound state to 0.95, 0.985, 0.99,

and 0.999 with our selection of TRANSFAC only redundant PWMs. We created the final selection of 102 TRANSFAC non-redundant PWMs with a prior probability of 0.999.

3.5.5 Shuffling PWMs

To investigate the influence of sequence specificity versus nucleotide content we also made copies of the selection of 102 PWMs, one with the last half of the MyoD or p53 PWM moved to the first half and one with a completely randomized order of the MyoD or p53 PWM nucleotides.

3.5.6 Binomial Tests

We performed binomial tests with the Math::Cephes module (*http://search.cpan.org/ dist/Math-Cephes/lib/Math/Cephes.pod*). These tests analyzed the number of TFBSs with a Sunflower probability greater than a user-defined flat cutoff (0.1 for PWM sets with mixed TRANSFAC/JASPAR PWMs and redundant TRANSFAC PWMs, and 0.01 for the larger selection of 102 TRANSFAC PWMs).

3.5.7 Identifying Potential Co-Regulators

To analyze if the best predicted PWM for each selection of PWMs represents TFs serving as potential co-regulators we converted the PWMs to factor names using the TRANSFAC website. These were used as input for Anni v2.0 (90). When a literature-based concept profile was not found for a factor name we used a gene alias that had a concept profile. Each set of concept profiles was then matched against the GO annotation consortium (91; 92) biological process concept file supplied in Anni.

3.5.8 Comparison to Existing Programs

We took the MyoD/Myog ChIP-chip promoter sequences (or gene IDs for PSCAN), all separated by CpG content, and ran these through CORE_TF and PSCAN. These were chosen since they all look for over-representation of TFBSs. For CORE_TF we used the same ChIP target sequences and 200 random promoters as used for the Sunflower analysis. PSCAN does not have the option to import sequences, but only a gene list. It also does not give a choice to give random data. We therefore gave input as Refseq IDs and defined sequences as close to the Sunflower work as possible: mouse promoters defined a -950 to +50 around the transcription start site. For all programs we used TRANSFAC PWMs (CORE_TF: a Match setting to minimize false positives in a non-redundant vertebrate TRANSFAC 11.2, PSCAN: TRANSFAC public).

3.6 Acknowledgements

We wish to thank Henk P.J. Buermans, Alison Meynert, Nicolas Rodriguez, Michel P. Villerius, and Maarten van Iterson for their input and technical support.

Funding was provided by Marie Curie Fellowship, the Center for Biomedical Genetics, the Netherlands, and the Center for Medical Systems Biology, co-funded by the Netherlands Genome Initiative.

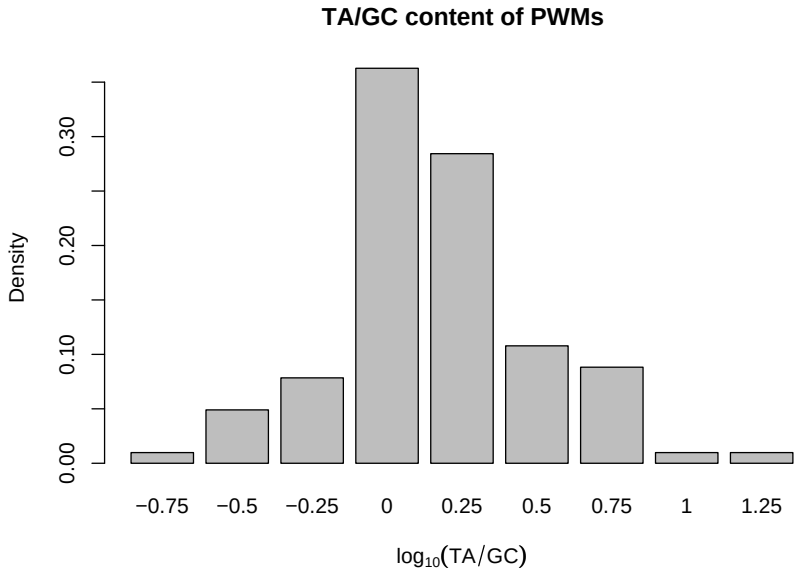
3.7 Additional Files

Additional File 3.1

Mixed*	TRANSFAC (small, redundant)		TRANSFAC (102, non-redundant)	
E2F1	V$AML1_Q6	V$MYOGENIN_Q6	V$AHRHIF_Q6	V$MYOD_Q6_01
ELK1	V$AP1_Q2_01	V$MYOGNF1_01	V$AIRE_02	V$MYOGNF1_01
Gata1	V$AP2ALPHA_02	V$MZF1_01	V$AP1_Q4_01	V$NF1_Q6_01
GATA2	V$COUPTF_Q6	V$MZF1_02	V$AP2ALPHA_01	V$NFAT_Q4_01
HLF	V$CREB_01	V$NERF_Q2	V$AP4_01	V$NFKB_Q6_01
MAX	V$CREL_01	V$NFKB_C	V$ARNT_01	V$NFY_Q6_01
MEF2A	V$E2F1_Q3	V$NFKB_Q6	V$ATF6_01	V$NKX3A_01
Myf	V$ELK1_01	V$NRF2_01	V$BACH2_01	V$NRSF_Q4
NFKB1	V$ELK1_02	V$P53_01	V$BCL6_Q3	V$OCT1_03
Pax6	V$FOXD3_01	V$PAX6_01	V$BRCA_01	V$OCT4_01
Pbx	V$FREAC2_01	V$PBX1_01	V$CEBP_Q3	V$P300_01
SOX9	V$FREAC3_01	V$PPARG_01	V$CREB_Q4_01	V$P53_02
SP1	V$FREAC7_01	V$PPARG_02	V$E2A_Q2	V$PAX2_02
SRF	V$GATA2_01	V$PPARG_03	V$E2F_Q6_01	V$PAX3_B
SRY	V$GATA3_01	V$PU1_Q6	V$E4BP4_01	V$PAX4_03
TBP	V$HAND1E47_01	V$RORA1_01	V$EGR1_01	V$PAX5_01
TP53	V$HEN1_01	V$RORA2_01	V$ER_Q6	V$PAX6_Q2
USF1	V$HFH3_01	V$RREB1_01	V$EVI1_03	V$PAX8_01
V$E2F1_Q3	V$HLF_01	V$SOX9_B1	V$FOXJ2_02	V$PBX1_03
V$ELK1_01	V$HNF1_Q6	V$SP1_01	V$FOXP1_01	V$POU1F1_Q6
V$GATA1_01	V$HOX13_01	V$SRF_01	V$FXR_Q3	V$POU3F2_02
V$GATA2_01	V$IRF1_01	V$SRY_01	V$GABP_B	V$POU6F1_01
V$HLF_01	V$IRF2_01	V$TAL1_Q6	V$GATA6_01	V$PPARA_02
V$MAX_01	V$MAX_01	V$TBP_01	V$GFI1_Q6	V$PPARG_01
V$MEF2_01	V$MEF2_01	V$TEF_Q6	V$GLI_Q2	V$RORA_Q4
V$MEF2_02	V$MYCMAX_01	V$USF_02	V$GRE_C	V$RSRFC4_Q2
V$MEF2_03	V$MYOD_01	V$VDR_Q3	V$HAND1E47_01	V$SF1_Q6_01
V$MEF2_04	V$MYOD_Q6	V$YY1_01	V$HEN1_01	V$SMAD3_Q6
V$MEF2_Q6_01	V$MYOD_Q6_01		V$HFH1_01	V$SOX9_B1
V$MYOD_01			V$HIC1_02	V$SP1_Q2_01
V$MYOD_Q6			V$HIF1_Q3	V$SP3_Q3
V$MYOD_Q6_01			V$HLF_01	V$SREBP_Q6
V$MYOGENIN_Q6			V$HNF1_Q6	V$SRF_Q6
V$MYOGNF1_01			V$HNF3ALPHA_Q6	V$SRY_02
V$NFKB_Q6			V$HNF3B_01	V$STAF_02
V$P53_01			V$HNF4_Q6_01	V$STAT1_01
V$P53_02			V$HNF6_Q6	V$TAL1BETAE47_01
V$PAX6_01			V$HOX13_01	V$TBP_Q6
V$PBX1_01			V$HOXA7_01	V$TBX5_01
V$SOX9_B1			V$IRF_Q6	V$TCF11_01
V$SP1_01			V$KAISO_01	V$TEL2_Q6
V$SRF_01			V$KROX_Q6	V$USF_Q6_01
V$SRY_01			V$LEF1_Q2_01	V$VDR_Q3
V$SRY_02			V$LHX3_01	V$VJUN_01
V$TBP_01			V$LXR_Q3	V$VMYB_02
V$TBP_Q6			V$MAF_Q6_01	V$WT1_Q6
V$USF_02			V$MAZ_Q6	V$YY1_Q6
V$YY1_01			V$MEF2_03	V$ZEC_01
YY1			V$MEIS1_01	V$ZF5_B
			V$MYB_Q3	V$ZIC2_01
			V$MYCMAX_03	V$ZNF219_01

Lists of all PWMs used in each PWM selection.
*TRANSFAC and JASPAR PWMs. TRANSFAC PWMs start with V$

Additional File 3.2



A density plot of the $\log_{10}$(TA/GC) of each PWM used in the selection of 102 PWMs.

Chapter 4

GAPSS: General Analysis Pipeline for Second-Generation Sequencers

Matthew S. Hestand[+1,2], Michiel van Galen[+2], Michel P. Villerius[1], Jaap W.F. van der Heijden[1], Gert-Jan B. van Ommen[1], Johan T. den Dunnen[1,2], Peter A.C. 't Hoen[1]

[1]The Center for Human and Clinical Genetics, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[2]Leiden Genome Technology Center, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[+]Equal contribution

not published

4.1 Abstract

Background: A simple to use generic system to perform primary analysis and annotation of second-generation sequencing data would be a valuable tool. Most software currently available is geared towards a specific application and requires considerable computer expertise.

Results: We have created GAPPS, which takes as input FASTA, FASTQ, or scarf files of second-generation sequencers' data and generates a report file (including the number of tags used as input and the number of tags aligned), UCSC genome browser tracks, files with basic annotation of regionally clustered tags, and a SNP report.

Conclusion: GAPSS is freely available, providing a simple to use tool for the average biologist to begin analysis of their second-generation sequencing data.

4.2 Background

Second-generation, also called next-generation, sequencing platforms (SSPs) can sequence gigabases of nucleotide sequence in a single run. Several platforms have been developed in the past years, each with their own unique qualities.

Processing and annotation of SSP data is difficult, requiring a basic level of bioinformatics expertise. This can include extensive knowledge of command line programming and difficult installations. This is often outside of the realm of the average biologist's knowledge. Often, for different applications, different analysis pipelines and programs were required. Chromatin immunoprecipitation coupled with SSP technology (ChIP-seq) analysis alone has had a multitude of applications developed for it, such as SISSRs (96), QuEST (97), a pipeline by Kharchenko and colleagues (98), and FindPeaks (99).

We focused on making a generic pipeline that can be used to perform a primary analysis of data from different SSPs and applications. Applications that can be addressed with this pipeline (with a reference genome) is analysis of SSP technology coupled with Cap Analysis of Gene Expression (CAGE) (29; 31), Serial Analysis of Gene Expression (SAGE) (28; 36; 100), and ChIP. It can also be used with basic SNP analysis compared to a reference genome. Our pipeline, titled GAPSS (General Analysis Pipeline for Second-generation Sequencers) automates primary SSP analysis in a user friendly manner.

4.3 Implementation

4.3.1 The Pipeline and Interface

GAPSS is controlled by a single Perl script that calls additional Perl scripts, Linux commands, and an alignment executable in a linear fashion (Figure 4.1), as described in the following sections.

GAPSS is run by executing a single script that prompts the user to answer several questions within a Linux terminal. For the faster version (discussed below) of GAPSS we also provide a GUI interface (Figure 4.2) programmed in PerlTk.

4.3.2 Sequence Editing

Step one (Figure 4.1) of GAPSS is to take all tags in each file and reduce them into a non-redundant set of tags. There is a user choice to retain the number of replicate tags or not, where replicate tags are considered to be derived from amplification of single products (101). Then, if requested, all tags are trimmed of their first nucleotide since this is often of low quality compared to other 5' nucleotides (102).

Linker sequences can potentially be in sequence reads when sequencing more cycles than the fragment length. Therefore, we provide the option to edit for defined linker sequences. These are removed from either the 5' or 3' ends of all tags, allowing for 0 or 1 mismatches. GAPSS tries to match the entire linker sequence first, and then shifts towards matching partial sequences nearer the requested end of the tag, requiring at least 3 nucleotides of linker sequence.

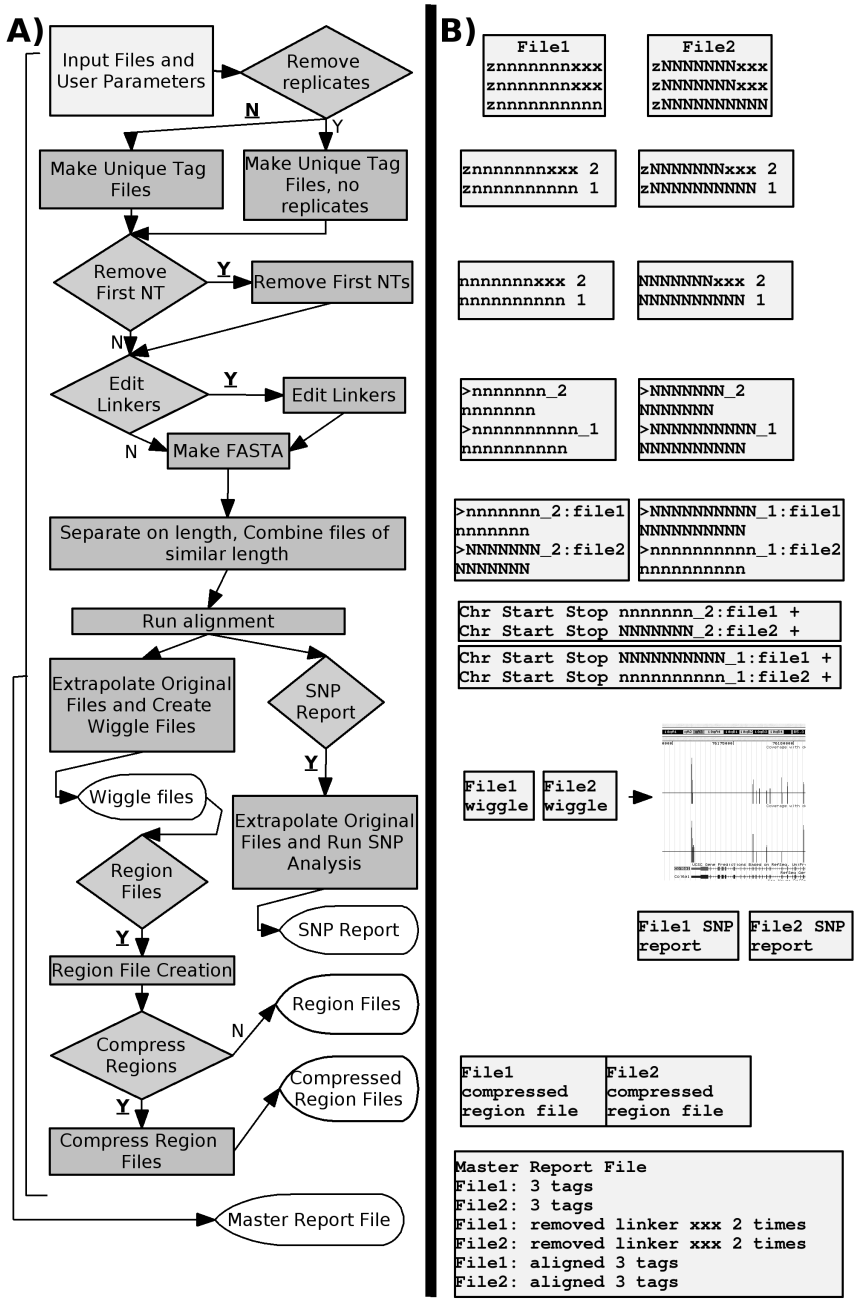


Figure 4.1: The GAPSS Pipeline: The data flow scheme (A) and arbitrary example files (B) for a GAPSS run. When a user option is available the example files are based on the choice presented in bold and underlined.

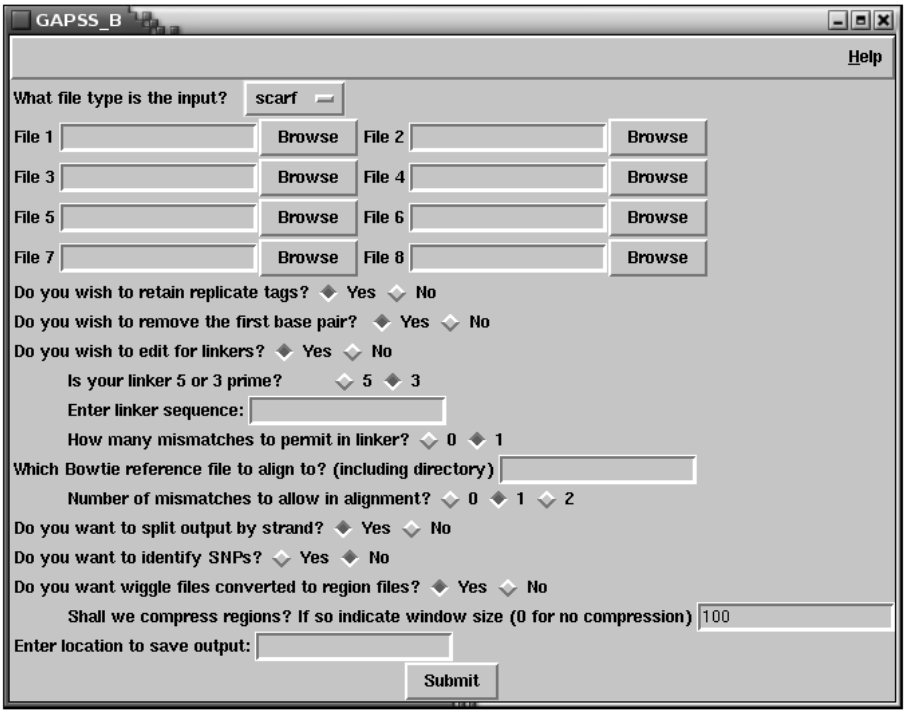


Figure 4.2: GAPSS_B GUI Interface

All sequences from every file are placed in FASTA files of unique sequence length due to constraints of some SSP alignment algorithms, including one which we utilize. We retain the names of files of origin in the FASTA headers.

4.3.3 Alignment

FASTA files containing specific length sequences are then run through the alignment tools Rmap (40) or Bowtie (42) (Figure 4.1). Both alignment tools are run for FASTA input on default parameters against a user defined reference, with a user choice in the number of mismatches permitted. For Bowtie we also implement the "–best" option to get the optimum alignment, not the first alignment encountered, for each sequence. All output files are then concatenated into one large file.

4.3.4 Wiggle and Region File Creation

These large alignment files of all concatenated data are separated back into individual files and converted to UCSC (103) style wiggle files, one file per original input file. This is possible since we retain file origins in our FASTA headers. There is also an option to export both DNA strands as one file or two separate files by strand. These can then be uploaded as "custom tracks" and viewed in the UCSC genome browser.

If requested, an additional step is entered to convert the wiggle files into region files. Region files are created by identifying all nucleotides that have adjacent hits

in the wiggle file. They include several columns of data, including region location (chromosome:start-stop), region length, the total number of tags hit on all nucleotides (similar to an "area-under-the-curve"), the average number of tags hit per nucleotide, the estimated number of tags in the region, the number of tags at the peak of the region, and the location of the peak of the region. Users can compress any number of regions within a user-defined window size into one region to suppress the presence of small gaps in the covered genomic sequence, retaining the afore mentioned region file data. These region files can serve as a post-GAPSS base for annotation (such as in Ensembl BioMart (3; 4)) and additional analysis.

4.3.5 SNP Report

In addition, GAPSS has the option to generate a SNP report (Figure 4.1). This is done by reading in the concatenated alignment outputs, sorting them by their file of origin, and extracting the location of mismatches in the sequence. Bowtie reports which nucleotides have mismatches, but for Rmap we infer this by comparing the aligned sequence back to the reference genome. All nucleotides with a mismatch are reported in a SNP report file that contains chromosomal position, the number of reads aligned to the reference, the number of reads aligned to each strand, the reference nucleotide, and the number of tags with an A, T, G, and C at this position.

4.4 Results and Discussion

4.4.1 Variants

Two variants of GAPSS have been created: GAPSS_R and GAPSS_B. GAPSS_R uses the alignment tool Rmap (40). GAPSS_B uses Bowtie (42) for alignment. Both have their advantages: Rmap for theoretical alignment accuracy and Bowtie for speed. Due to long run times GAPSS_R is only implemented as a command line executable, whereas GAPSS_B has been implemented as both a command line and GUI interface (Figure 4.2).

4.4.2 Usage

GAPPS is run by executing a Perl script that enables a command line or GUI interface. Users answer several questions and GAPSS then automates the entire analysis process.

This takes as input FASTA, FASTQ, or scarf (Illumina Genome Analyzer's pipeline GERALD output) format files and converts them to a variety of output: a general report file, wiggle files (viewable as tracks in the UCSC Genome Browser), region files, and a SNP report. The report file contains information on the run, including the number of tags in each input file, the number of tags aligned, and additional details on sequence analysis and editing.

We have successfully tested GAPSS on a variety of Illumina Genome Analyzer and Roche 454 data. With the option to use FASTA and FASTQ format input we believe it can also be used with additional platforms.

GAPSS is run on Linux. For Ubuntu users, an install script is included to easily install GAPSS_B and additional files. For other systems and GAPSS_R a manual

installation is available. The individual Perl scripts are also available for bioinformaticians to tailor make their own pipelines.

4.4.3 Performance

Using GAPSS_B, with 4xCPU(W5580 @ 3.20 GHz), data from 2 experiments (one human ChIP-seq experiment, one mouse DeepCAGE experiment, both with 2 lanes of data from the Illumina Genome Analyzer) was analyzed against reference genomes in approximately 70 minutes per experiment using approximately 2 GB memory (Additional File 4.1).

4.4.4 Plans

GAPSS has been programmed in a very modular fashion so we may incorporate newer software, such as improved alignment programs, as technology improves. As hardware and software improve the speed of analysis will improve, hopefully allowing for a web-based GAPSS in the future. This would enable even easier access and usage to the average biologist.

4.5 Conclusions

GAPSS is a simple to run generic pipeline, providing biologists with a comprehensive system to begin analyzing their SSP results. GAPSS and example data is freely available for download at *www.lgtc.nl/GAPSS*.

4.6 Availability and requirements

Project name: GAPSS
Project home page: *www.lgtc.nl/GAPSS*
Operating system(s): Linux
Programming language: Perl
Other requirements: Rmap and BioPerl for GAPSS_R. Bowtie and PerlTk for GAPSS_B.
License: GNU General Public License
Any restrictions to use by non-academics: none

4.7 Authors Contributions

MH was involved in developing the concept, primary programming, debugging, and manuscript drafting. MG performed primary programming and debugging. MV performed GUI programming, debugging, and installation assistance. JH performed SNP programming and debugging. GO, JD, and PH were involved in developing the concept and manuscript drafting.

4.8 Acknowledgements

We wish to thank Ivo Fokkema for computational assistance and Yavuz Ariyurek for his Illumina Genome Analyzer expertise. This project was supported by grants from the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) and the Center for Biomedical Genetics (in the Netherlands).

4.9 Additional Files

Additional File 4.1
Performance Evaluation details

Maximum Number of CPUs Used: 4 x 3.40GHz
Available Memory: 32GB
GAPSS version used: GAPSS_B (GUI)
Settings used across runs:
-file type: scarf
-retain replicate tags
-remove first nucleotide from tags
-allow 2 mismatches when aligning to reference genome
-create SNP reports
-create region files
-compress region files (size 100)
Reference files were obtained from the Bowtie website
(*http://bowtie-bio.sourceforge.net/tutorial.shtml*)

Test Data 1: 2 Human ChIP-seq samples
(one lane of the Illumina Genome Analyzer per sample)
-read length: 32 NT
-11826172 total tags
Test Data 2: 2 Mouse Deep-Cage samples
(one lane of the Illumina Genome Analyzer per sample)
-read length: 36 NT
-9946382 total tags

Test Data	Reference	Edit linkers (NT long, mismatches)	Separate by strand	Memory Used	Approx. Run Time
1	Human (contigs, 36)	No	No	7% (∼2.24GB)	144 minutes
2	Mouse (contigs, 37)	Yes (21, 1)	Yes	20% (∼6.4GB)	109 minutes

Chapter 5

Genome-Wide Assessment of Differential Roles for p300 and CBP in Transcription Regulation

Yolande F.M. Ramos[1†], Matthew S. Hestand[2,3†], Matty Verlaan[1], Elise Krabbendam[1], Yavuz Ariyurek[2], Michiel van Galen[2], Hans van Dam[1], Gert-Jan B. van Ommen[3], Johan T. den Dunnen[2,3], Alt Zantema[1‡], Peter A.C. 't Hoen[3‡]

[1]Department of Molecular Cell Biology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.
[2]Leiden Genome Technology Center, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[3]Department of Human and Clinical Genetics, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[†]Both authors contributed equally to the work presented.
[‡]Both authors contributed equally to the work presented

Nucleic Acids Res. 2010 Apr 30. [Epub ahead of print]
Parts of this manuscript have been adapted to more appropriately fit this thesis.

5.1 Abstract

Despite high levels of homology, transcription coactivators p300 and CREB binding protein (CBP) are both indispensable during embryogenesis. They are known to largely regulate the same genes. To identify genes preferentially regulated by p300 or CBP, we performed an extensive genome-wide survey using ChIP-seq on cell-cycle synchronized cells. We found that 57% of the tags were within genes or proximal promoters, with an overall preference for binding to transcription start and end sites. The heterogeneous binding patterns possibly reflect the divergent roles of CBP and p300 in transcriptional regulation. Most of the 16,103 genes were bound by both CBP and p300. However, after stimulation 89 and 1944 genes were preferentially bound by CBP or p300, respectively. Target genes were found to be primarily involved in the regulation of metabolic and developmental processes, and transcription, with CBP showing a stronger preference than p300 for genes active in negative regulation of transcription. Analysis of transcription factor binding sites suggest that CBP and p300 have many partners in common, but AP-1 and Serum Response Factor (SRF) appear to be more prominent in CBP-specific sequences, whereas AP-2 and SP1 are enriched in p300-specific targets. Taken together, our findings further elucidate the distinct roles of coactivators p300 and CBP in transcriptional regulation.

5.2 Introduction

The primary mechanism to control cellular processes, such as proliferation and differentiation, is by regulation of gene expression (reviewed in (104; 105; 106)). Gene expression is a highly coordinated process that results in the synthesis of messenger RNA after recruitment of histone modifying factors, the pre-initiation complex, and transcription factors (TFs) to regulatory regions of the chromatin. The histone modifications that take place during this process, including methylation and acetylation, play a critical role in gene regulation, and defects have been implicated in many pathological conditions from cancer to autoimmune diseases (107; 108; 109). Recently, chromatin immunoprecipitation (ChIP) has been extensively applied in combination with high-throughput sequencing to map genome-wide chromatin modification profiles in human T cells (110; 111) and in mouse ES cells (112). Binding sites of the insulator binding protein CTCF (110), RNA pol II (113; 110) and several TFs (114; 115; 116; 35) have also been mapped. The acetylation profile in primary human T cells was further investigated by determining the binding of several histone deacetylases (117) and histone acetyltransferases (HATs) including p300. Binding of p300 was found both at genes and at intergenic DNase hypersensitive sites, consistent with binding to enhancers, found in other p300 ChIP-sequencing experiments (118; 119).

The HAT p300 and its family member CREB-binding protein (CBP) are transcription coactivators for a broad range of genes involved in multiple cellular processes such as proliferation, differentiation, apoptosis, and DNA repair (reviewed in (120; 121)). In addition, a number of studies suggested the involvement of p300 and CBP in pathological disorders such as the RubinsteinTaybi Syndrome (reviewed in (122)) and the development of cancer (reviewed in (123)). Originally, CBP was identified through its association with the phosphorylated TF CREB (124), but CBP and p300 also interact with many other TFs, such as cJun (125), p53 (19), and MyoD (126). Apart from the transcriptional regulation through acetylation of histones and other factors, p300 and CBP can also act as a bridge or as a scaffold between upstream TFs and the basal transcription machinery.

A crucial role for both p300 and CBP in development was shown in mice with a homozygous deletion of either gene (*Ep300* and *Crebbp* for the proteins p300 and CBP) resulting in embryonic lethality at a very early stage (20; 21). Interestingly, the double heterozygous $Ep300^{+/-}/Crebbp^{+/-}$ mice also die *in utero* (20), indicating that a fine-tuned balance in the expression of both proteins is needed to ensure the normal development. From phenotypic changes in the knock-out mice it is indicated that p300 and CBP have different functions, which has been further illustrated in additional *in vivo* studies (127; 128; 129). A comparison between the acetyltransferase domains of p300 and CBP showed that they differ structurally (130). In part, this might contribute to their functional differences. However, the current detailed mechanism of action of p300 and CBP and the differences between these transcription coactivators is not clear.

In contrast to the *in vivo* situation, most studies with cells cultured *ex vivo* show similar functions for p300 and CBP, and only limited differential roles for p300 and CBP have been described (reviewed in (120)). To obtain a better insight into genes regulated by the general transcription coactivators p300 or CBP next-generation sequencing of ChIP genomic fragments (ChIP-seq) (35) was performed. ChIP-seq and

ChIP-on-microarray (ChIP-chip) have high correspondence in results, but ChIP-seq offers the advantages of requiring less input material, potential to identify binding sites with low affinity, not being limited to target regions (*i.e.* probes on a microarray), not having hybridization errors and it is less costly for whole genome analysis (35). In this study, we used the glioblastoma cell line T98G. T98G cells can easily be synchronized by serum-deprivation and reintroduced into the cell cycle upon stimulation with serum and TPA. Previously, RNA pol II ChIP was performed in growth factor stimulated T98G cells (131), and this showed that 30 minutes upon growth factor stimulation occupancy of the polymerase at the promoters of immediate early genes was maximal. We observed that maximal occupancy of p300 and CBP at promoters of immediate early genes was also around 30 min (Y.F.M.R., unpublished results).

We show p300 and CBP binding to the chromatin in quiescent and stimulated cells, and alterations in their binding to a large number of genes after stimulation. In most cases there is overlap between regions bound by p300 and CBP, but we also identified distinct regions of binding, indicating specific targets for each of these acetyltransferases. Bound regions were analyzed genome-wide for their position relative to genes and were found to have a preference for transcription start sites (TSSs) and transcript ends. Interestingly, functional classification of target genes suggests that CBP is more involved in the regulation of transcription inhibition than p300. A list of TFs that might be involved in the transcription regulation of the identified genes together with p300 and/or CBP was obtained by searching for enriched TF binding sites (TFBSs) in the bound regions. Results show previously established binding partners, and suggest differences for p300 and CBP in their preferences for TFs.

5.3 Materials and Methods

5.3.1 Cell Culture, ChIP, qPCR, and Sequencing

Human glioblastoma T98G cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), penicillin (100 μg/ml) and streptomycin (100 μg/ml). Prior to stimulation with serum (20%) and tetradecanoyl phorbol acetate (TPA 100 ng/ml; Sigma), cells were serum starved for 23 days (DMEM supplemented with 0.1% FBS).

For sequencing, chromatin was isolated from serum-starved cells (T0) and from cells stimulated for 30 min with serum and TPA (T30). Chromatin from T30 samples was prepared in duplicate, each being used for individual ChIPs, sequencing and downstream analysis. In addition, for more time-point specific data (analyzed only by ChIP and quantitative PCR) we isolated chromatin at 0, 2, 5, 15, 30, 60, 90, 120 and 360 min following stimulation. Chromatin was prepared and ChIPs were carried out as previously described, including fragmentation by sonication (132) (fragment size 500 bp). Immunoprecipitations were performed for p300 using the p300-(2) antibody produced in our lab (125), and for CBP with a commercially available antibody (A22 from Santa Cruz).

For Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) analysis, RNA was isolated using the SV Total RNA isolation System (Promega Corporation Benelux), according to the manufacturers' protocol, and first-strand cDNA synthesis was performed using 1 μg of RNA and ImProm II reverse transcriptase (Promega Corporation

Benelux).

Quantitative PCR for ChIP and for cDNA samples was carried out using the Applied Biosystems 7900HT Fast Real-Time PCR System with SYBR Green PCR Master Mix (Applied Biosystems Europe). Primers were designed using the Primer Express program from Applied Biosystems (for sequences of primers see Additional Table 5.1). Efficiency of the ChIP is presented as percentage of the input. Expression levels of the genes as determined by quantitative RT-PCR were normalized to *GAPDH*, and fold induction was calculated with reference to the untreated samples (t = 0 minutes).

For ChIP-seq all samples were prepared with Illumina's DNA sampleprep Kit (FC-102-1001) according to the manufacturer's protocol. Single ends of each sample were then sequenced on a single lane of the Illumina Genome Analyzer (GAI for samples CBP T0 and T30-1 and p300 T0 and T30-1, GAII for samples CBP T30-2 and p300 T30-2) for 36 cycles.

Illumina Genome Analyzer Sequencing Analysis

Sequencing results were run through the standard Illumina GAPipeline (v1.0 for GAI runs and v1.3 for GAII runs) to convert images to reads (unaligned sequences produced by the Illumina Genome Analyzer) and edit for quality (FIRECREST, Bustard and GERALD). A general overview of the entire ChIP-seq analysis is provided in Additional Figure 5.1A. The reads were then trimmed to the first 32 bp to remove lower quality base calls at the 3′-end of the read. These were then run through the developing GAPSS_R (*www.lgtc.nl/GAPSS*) pipeline. This pipeline took the reads, removed the first base pair (often low quality compared to other 5′ nucleotides), converted to FASTA format, aligned to the human reference genome (NCBI build 36) with Rmap v0.41 (40), permitting up to two mismatches, and exported tags (the term for aligned reads) into region files (merging adjacent nucleotides with at least one aligned read into one region, followed by compressing those regions within 100 bp into one (based on a range of compression sizes, see below and Additional Figures 5.1B and 5.2). The pipeline also created wiggle files (viewable in the UCSC genome browser (103)). These tracks had positions with only a single read removed, in order to create more manageable files.

All unedited wiggle files were concatenated to one with custom Perl scripts and converted to a region file (a range of compression windows (20, 50, 100, 150 and 200 bp) were used) with GAPSS_R scripts. The compression windows account for small gaps in the genomic sequences covered, such as the result of non-unique genomic sequences (Rmap does not map to these). An appropriate compression size is hard to determine, considering a bigger window results in less regions (Additional Figure 5.2) and therefore specificity, but covers larger genomic repeats. We settled on a window of 100 bp to retain a large number of regions, while at the same time accounting for small repetitive elements. This consensus region file had the number of tags from the individual region files mapped to it with a custom Perl script. To make data more manageable and reduce background or very low affinity binding we removed regions with <6 tags (total over all samples). To further reduce the noise only regions with at least 1 tag/million reads aligned (18.1 tags across all total samples) were evaluated. Without applying this threshold performance was poorer, as addressed in the results.

To annotate regions we downloaded from Ensembl 54 Biomart (3; 4) for all genes (with an HGNC ID) the chromosomal location, strand, gene start, gene end, transcript start, transcript end and gene ID. These were loaded into a custom mysql database that was queried to annotate regions for overlap with genes (including flanking 1 kb). We also annotated for distance to the nearest TSSs and transcript ends. Histograms were plotted in the statistical language R to visualize the distance to TSSs and transcript ends.

5.3.2 Statistical Analysis

The statistical language R was used to evaluate reproducibility and overlap across samples and to determine genes with differential TF binding across different conditions. To be able to compare data across samples, samples were scaled to the average total number of tags per condition/coactivator. A square root transformation was applied before calculating the reproducibility and comparability across samples. This was to stabilize the variance, inherent to the counting process, over the entire intensity range (133), and to spread the data points better over the intensity range (Additional Figure 5.3). After this, to give a better estimation of the comparability of the data from the different samples Pearsons correlations were calculated in R. This was done on all regions with abundance >1 tag/million tags and a square root transformation applied before calculating the correlations. The Pearsons correlations on the linear scale were slightly lower.

Subsequently, data were summarized at the gene level by adding all tags within a gene or its 1 kb flanking regions. To determine the genes different between conditions/coactivators Fishers exact p-values were calculated in R. For each individual gene, a two-by-two table was created containing the number of tags for this gene in condition 1 and condition 2 and the total number of tags in condition 1 and condition 2. We then applied the method of Benjamini and Hochberg to correct for multiple testing.

5.3.3 Functional Classification

A list of 250 genes, identified as most significantly different between the time points for each coactivator (T30/T0 with adjusted p-value <0.001), was uploaded in DAVID 2008 (89; 134) for functional enrichment analysis. To obtain a general impression of the types of processes in which CBP and p300 are involved, functional annotation charts were generated for the Gene Ontology (GO) term GOTERM_BP_ALL (91; 92) using a human background.

In addition, significantly different genes at T30 were divided into two groups where either CBP or p300 binding was higher. From these groups, the 250 genes most significantly different were uploaded in DAVID 2008 for functional enrichment analysis. Individual GO-terms with a p-value <0.001 are shown for genes with higher CBP or p300 binding.

5.3.4 CORE_TF Analysis for TF Partners of p300/CBP

We took the same significant gene sets as from the functional analysis and retrieved the most substantially sequenced region (most number of tags in this particular re-

gion) for these genes. These regions were extended at both sides to a final length of 2 kb and sequences retrieved with Ensembl Perl API. As a background set, we retrieved 3000 random genes' TSSs from Ensembl Biomart that were located on chromosomes 1−22, X and Y and retrieved the sequences +/−1 kb from these TSSs. The regions based on significantly different genes were entered into CORE_TF (76) as experimental sequences and the random TSS sequences were entered as background regions. We evaluated enrichment of TFBSs (defined as TRANSFAC (51) position weight matrices) in the experimental sequences using the most stringent Match setting (55; 51) to minimize false positives. P-values representing the significance of over-representation were calculated with a binomial test.

5.4 Results

5.4.1 Initial Sequencing Analysis

Stringent regulation of gene expression is fundamental to control cellular processes such as proliferation and differentiation. The general coactivators p300 and CBP play an important role in the regulation of gene transcription by virtue of their acetyltransferase activity. We set out to determine and compare genes regulated by p300 and CBP. Chromatin was isolated from serum-starved (T0) and from stimulated (T30, done in duplo) human glioblastoma cells and ChIP-seq performed using CBP-and p300-specific antibodies.

Sequence files generated by the Illumina GAPipeline were submitted to the NCBI Sequence Read Archive (SRA; *http://www.ncbi.nlm.nih.gov/Traces/sra*: SRS009476, SRS009457, SRS009477, SRS009478, SRS009479 and SRS009480) (135). The reads passing quality control were mapped to the human reference genome and adjacent tags were joined into regions (Table 5.1). We also have made sequencing data available as UCSC hg18 viewable wiggle tracks (excluding positions with only one tag aligned, Additional Files 5.1-5.6).

5.4.2 Preferential Binding in Genes and Promoters

Without applying a threshold of 1 tag/million tags, we found low overlap of identified regions in the replicated samples indicating that regions with low abundance represent noise (data not shown). With the threshold of 1 tag/million tags, we found a high consistency in the identified regions between all samples (47.96 and 47.43% overlap between CBP and p300 at T0 and T30, respectively; Table 5.2). Concordantly, the reproducibility between biological replicates was high (Pearsons correlation: 0.77 and 0.87 for CBP and p300, respectively). A similarly high correlation was found across the different samples (Pearsons correlation 0.81 on average between all time points and coactivators; Additional Table 5.2), indicating relatively minor differences in the distribution of p300 and CBP binding sites across the genome. In subsequent analyses, datasets of the T30 biological replicates were summed and treated as one sample, which provided us with high-quality results.

To study the biological implications of our data, we annotated the regions obtained from sequencing with Ensembl and found that the sequenced regions covered 16,103 annotated genes in total. When looking at conditions and coactivators independently

Table 5.1: Sequencing Results

		CBP		
t(min)	# reads	# aligned	% aligned	# regions*
T0	5498759	4018590	73	713141
T30[1]	6389605	4849826	76	889781
T30[2]	6047530	5001204	83	851988
		p300		
t(min)	# reads	# aligned	% aligned	# regions*
T0	6327413	5086340	80	841029
T30[1]	6446269	5156450	80	802627
T30[2]	6065594	5124836	84	684222

The total number of reads, reads aligned, percentage aligned and number of regions created (*after compressing regions within 100 bp into one and excluding regions composed of only a single tag) for each condition (T0: quiescent cells and T30: 30 min after growth factor stimulation) and for each transcription coactivator (CBP or p300). For T30-independent biological replicates were sequenced as indicated by [1] and [2].

Table 5.2: Region overlap

	CBP T0	CBP T30[2]	CBP T30[1]	p300 T0	p300 T30[2]	p300 T30[1]
CBP T0	267562	129089	133493	140245	120092	134799
CBP T30[2]		315020	143160	152520	136405	148669
CBP T30[1]			322556	151519	136685	150180
p300 T0				322354	138933	158708
p300 T30[2]					267804	139592
p300 T30[1]						304880

The number of regions, after applying thresholds (> 1 tag/million tags), overlapping between conditions (T0 and T30) and coactivators (CBP and p300). For T30-independent biological replicates were sequenced as indicated by [1] and [2].

there were 16,045, 16,075, 15,684, and 15,996 genes identified as bound by CBP at T0 and T30, and by p300 at T0 and T30, respectively. We observed similar percentages of tags in genes and their 1 kb flanking regions in all samples (57.08, 57.10, 57.30 and 59.93% for CBP-T0, CBP-T30, p300-T0 and p300-T30, respectively). Therefore, both CBP and p300 appear to be needed to maintain basal levels of expression in quiescent cells as well as to activate or repress transcription after serum stimulation.

Previous studies have focused on the binding of p300 to enhancers (118; 117). First, we evaluated the distance for all regions bound by CBP or p300 to the nearest TSS and transcript end (polyadenylation site). We found that genome-wide, 57% of all tags could be annotated to genes (1 kb) and a clear preference for TSSs and transcript ends was observed (Figure 5.1A and B). There were no apparent differences between the profiles of CBP and p300 (data not shown). Also, different from what has been shown before for most TFs or histone modification maps, p300 and CBP show three distinct patterns of binding, including a distinguished peak (binding to a specific site like the TSS, *e.g. ZNF688*; Figure 5.1C), binding across the gene with

no clear preference for a specific region (*e.g. EGR1*; Figure 5.1D), and so-called U-shaped binding (binding across the gene with a bias toward the TSS and transcript end, *e.g. DUSP1*; Figure 5.1E).

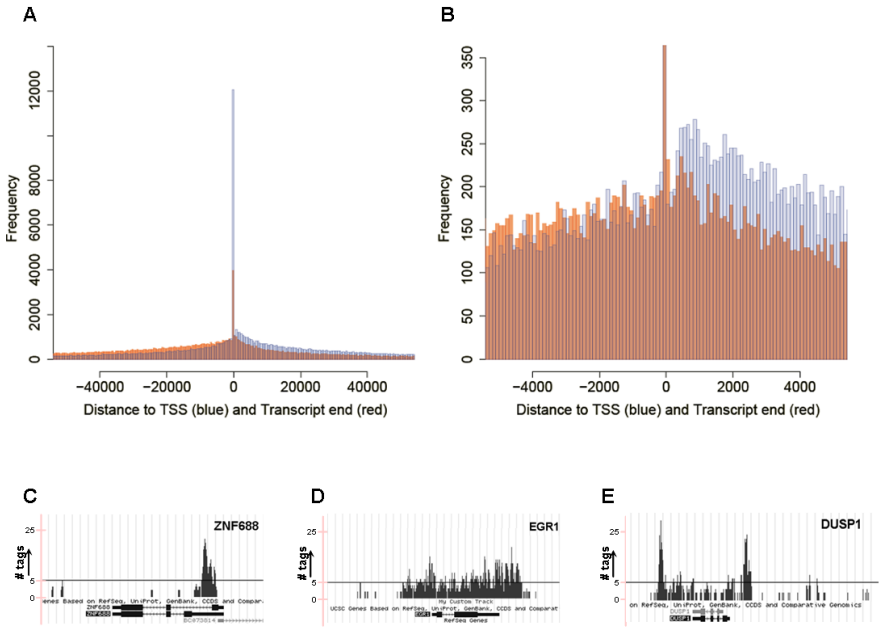


Figure 5.1: Histogram for the compilation of ChIP-seq regions showing the frequency of the distance from the localization of a sequenced region to the nearest transcription start site (blue) and transcript end (red) (full plot in (A), zoomed in (B)), which indicates a preference for binding to TSSs and transcript ends (color figure available at *http://nar.oxfordjournals.org/cgi/content-nw/full/gkq184v1/F1*). Representative examples of the different types of binding are shown as custom tracks on the UCSC genome browser: binding to a specific site resulting in a peak (C), binding across the gene (D), and U-shaped binding, with binding across the gene with preference for both TSS and transcript end (E). The y-axis indicates the number of tags aligned at each position in the genome. The black line in Figure 5.1C-E indicates a value of 5 tags in the custom tracks.

5.4.3 Differential Binding by CBP and p300

With most data corresponding to a genic region, we focused our following analyses to genes, and on those regions within 1 kb upstream of TSSs and 1 kb downstream transcript ends, (16,103 genes across all four samples). Since we were especially interested in genes that were preferentially regulated by p300 or CBP during entry in the cell cycle, a Fishers exact test was performed to determine statistically significant differences in the total number of tags localized to a certain gene in different conditions

(between time points or between coactivators) studied.

Despite high overlap in regions bound by CBP and p300 in quiescent and in stimulated cells (Table 5.2), there was also a considerable number of quantitative changes in CBP and p300 binding upon stimulation. Significant differences between p300 and CBP binding was found for 120 and for 1611 genes at a false discovery rate of 0.1% at T0 and T30, respectively (Figure 5.2A). At a false discovery rate of 1% this was 256 and 2502 at T0 and T30, respectively (Additional Table 5.3). From the genes differentially bound by p300 and CBP in quiescent cells (T0), only 25 did not have significantly different binding upon growth factor stimulation (Figure 5.2A). These results indicate very high overlap in genes bound by p300 as well as by CBP in the quiescent state and a divergence of the roles of CBP and p300 mainly during periods of activated transcription. Analysis of the 250 genes that were most significantly different in our data, showed that for the majority p300 binding was higher than CBP binding (191 and 227 of 250 genes, for T0 and T30, respectively).

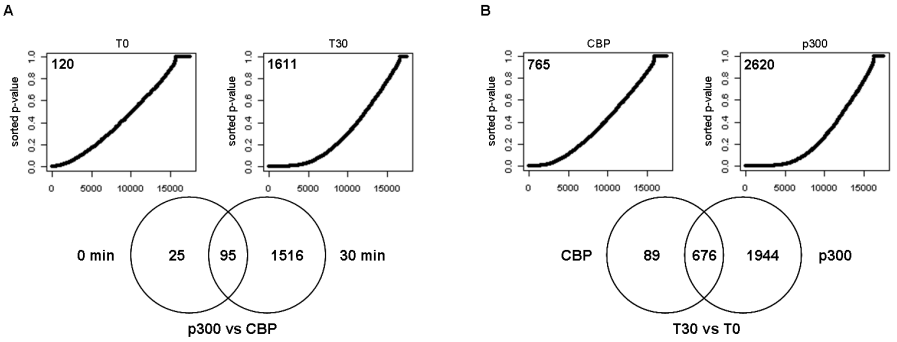


Figure 5.2: Genes differentially bound by CBP and p300 (A) and between time points (B). P-values (Fishers exact test) for the indicated comparisons were sorted in rising order and plotted (Upper panels). Under the null hypothesis of no significant differences, this would give a straight line on the diagonal. However, as becomes evident by the curve shape there is a bias towards low p-values. The number of genes with significant differences between conditions are indicated in the graphs (false discovery rate of 0.1%). Venn diagrams (Lower panels) demonstrate the number of significantly different bound genes, as shown in the plots above, that overlap between time points (A) or coactivators (B).

When comparing between time points, we found 765 genes differentially bound by CBP and 2620 genes differentially bound by p300 (Figure 5.2B). Of the 250 genes, which were most significantly different between time points, the majority (209 and 155 for CBP and p300, respectively) demonstrated higher binding for both, p300 and CBP, at T30 when compared to T0. In addition, the majority of genes with changed binding of CBP after stimulation also demonstrated difference in binding by p300 (676 out of 765 and 2620 genes, respectively; Figure 5.2B). The apparently higher number of genes with significant changes in p300 binding is likely due to the higher efficiency of the p300 antibody causing better signal-to-noise ratios and higher sensitivity in the detection of quantitative changes in binding profiles (see below). The high level of

overlap between coactivators can explain the restricted number of differences found thus far in functions of p300 and CBP.

We present a full list of genes bound by CBP and p300 at T0 and T30 in Additional File 5.7. Table 5.3 lists the 10 genes for which levels of binding differ most significantly between the four samples. Among the genes with strongest CBP and p300 binding and most significantly different between T30 and T0 are many immediate-early genes that are bound by both p300 and CBP (Table 3; *e.g. ATF3*, *FOSB*, and *DUSP1*).

Table 5.3:

CBP T30 vs T0			p300 T30 vs T0		
gene	p-value	ratio	gene	p-value	ratio
CATSPER3	0	2.58	THBS1	6.13×10^{-251}	6.07
ATF3	5.58×10^{-101}	4.51	ATF3	1.50×10^{-240}	5.83
TRIP13	4.63×10^{-94}	11.07	FOSB	2.22×10^{-213}	14.03
CYR61	4.15×10^{-80}	6.43	CYR61	1.27×10^{-187}	6.33
FOSB	5.04×10^{-75}	10.03	EGR1	1.66×10^{-178}	14.11
SMAD3	8.18×10^{-72}	2.09	TPM1	2.25×10^{-175}	4.81
TMEM49	4.60×10^{-71}	2.32	DUSP1	2.05×10^{-147}	6.81
MYH9	2.15×10^{-69}	2.39	MYH9	8.27×10^{-144}	2.84
CRISPLD2	1.45×10^{-67}	2.89	NR4A1	4.74×10^{-143}	13.16
THBS1	1.32×10^{-66}	4.15	CRISPLD2	1.45×10^{-140}	4.04
T0 p300 vs CBP			T30 p300 vs CBP		
gene	p-value	ratio	gene	p-value	ratio
CXXC1	8.28×10^{-229}	22.71	CXXC1	3.06×10^{-43}	20.32
AKT1S1	1.39×10^{-192}	6.03	MKKS	1.66×10^{-41}	3.87
FBXL19	9.96×10^{-172}	5.56	CATSPER3	1.07×10^{-30}	1.44
MKKS	1.67×10^{-154}	4.73	AKT1S1	6.27×10^{-24}	4.46
C3orf19	2.39×10^{-135}	7.93	FAM40A	1.94×10^{-22}	4.21
BSCL2	3.52×10^{-130}	8.25	FBXL19	1.09×10^{-21}	3.27
THBS1	5.46×10^{-120}	2.1	ZNF350	1.09×10^{-21}	9.14
MADCAM1	1.24×10^{-112}	7.85	METTL3	1.50×10^{-19}	13.47
ZNF175	1.01×10^{-103}	17.86	MADCAM1	1.30×10^{-18}	7.2
C1orf174	1.17×10^{-102}	9.84	C1orf174	1.84×10^{-17}	7.25

Top 10 genes that are most significantly different between time points and coactivators according to the p-values of the Fishers exact test. The ratio shows the quantitative difference in binding as expressed by the number of tags between the two samples that are compared: from T30 and T0 (upper half of the table) and from p300 and CBP (lower half of the table).

5.4.4 Validation

To validate our results and to refine the temporal resolution of the experiment, genes were selected to further characterize with ChIP and quantitative PCR in a time-course from 0 to 360 min following stimulation with serum and TPA. The genes included genes bound by both CBP and p300 and genes unique to one of the coactivators, and spanned a wide range of significance values (Figure 5.3E). In general, the recov-

ery obtained (as a percentage of the input) for CBP is lower than for p300 (Figure 5.3A-D), consistent with the generally lower number of tags for CBP in each region of the ChIP-seq experiment (significant quantitative correlation between results of the qPCR and ChIP-seq experiment for the genes presented here are shown in Additional Figure 5.4). The qPCR results also confirm the differential binding across time points established with ChIP-seq analysis for all genes analyzed (Figure 5.3A-D and Additional Figure 5.5), and demonstrate that for most genes the temporal binding pattern is comparable between CBP (black bars) and p300 (white bars). This is true for the increased binding to the promoter of *CTGF*, as well as for the decreased binding to the promoter of *ZNF608* in stimulated cells compared to unstimulated cells (Figure 5.3A and B). Binding to the promoter of *CDK5* differs for p300 and CBP (Figure 5.3C). Binding of p300 is increased in time with a maximum at 60 min post-stimulation, while there is hardly any change in the binding of CBP. These results correlate with the statistical analysis, that demonstrated significant changes between p300 and CBP at T30, and a significant increase in p300 but not CBP binding between T30 and T0 (Figure 5.3E). The binding of p300 and CBP to the *SERPINE1* gene increased significantly over time (p-value of 1.88×10^{-17} for CBP T30 versus T0 and 1.59×10^{-24} for p300 T30 versus T0). Inspection of the wiggle track (Figure 5.3D) revealed that p300 and CBP bound mainly to the 3′-UTR and to a lesser extent to the region around the TSS of the *SERPINE1* gene. Also, the small increase observed around the TSS could be confirmed by qPCR. The wiggle file for *SERPINE1* also shows a stronger binding >2 kb upstream of the TSS (Figure 5.3D). The interaction to this putative enhancer region and the change upon stimulation was also confirmed by qPCR of ChIP samples (Additional Figure 5.5J).

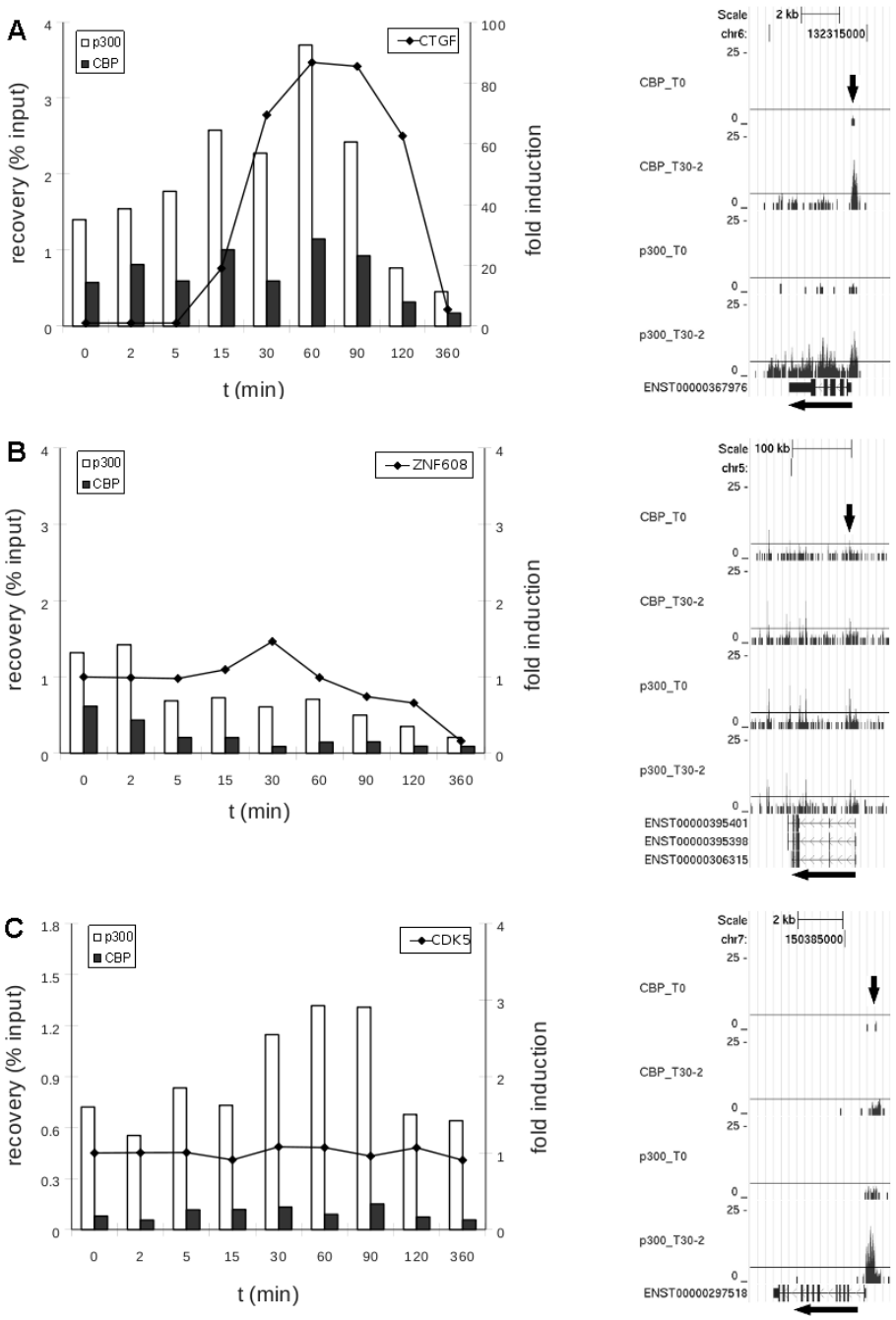
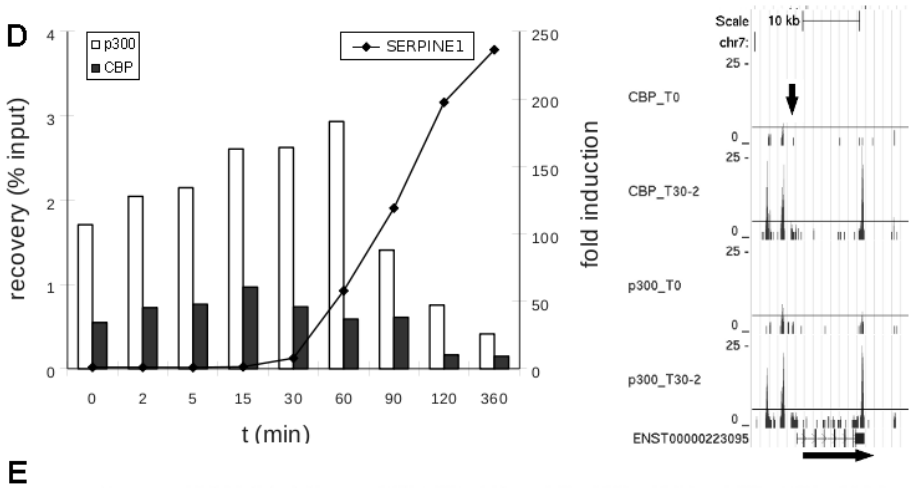


Figure 5.3 (continues on next page)

To evaluate whether changes in p300 and CBP binding also affected gene expression, we performed quantitative RT-PCR for the genes *CTGF, ZNF608, CDK5,* and *SERPINE1* (Figure 5.3A-D: the line in the graphs shows fold induction in the time course). For the three genes with increased binding, two (*CTGF* and *SERPINE1*)



E

Gene	CBP_T30vsT0		p300_T30vsT0		T0_p300vsCBP		T30_p300vsCBP	
	adj.p-value	ratio	adj.p-value	ratio	adj.p-value	ratio	adj.p-value	ratio
CTGF	1.23×10^{-23}	5.24	1.01×10^{-100}	8.95	5.22×10^{-1}	1.47	8.06×10^{-56}	2.51
ZNF608	7.50×10^{-5}	0.81	1.03×10^{-60}	0.55	8.37×10^{-10}	1.37	8.69×10^{-2}	0.93
CDK5	5.48×10^{-1}	1.27	7.78×10^{-5}	1.82	8.80×10^{-1}	1.16	2.81×10^{-5}	1.66
SERPINE1	1.88×10^{-17}	3.60	1.59×10^{-24}	3.57	1.00	1.03	9.29×10^{-1}	1.02

Figure 5.3: (continued from previous page): ChIP-analysis for time-course experiment (0, 2, 5, 15, 30, 60, 90, 120 and 360 min after stimulation of serum-starved T98G cells with serum and TPA). Shown are graphs for qPCR results (x-axis: time in minutes; left y-axis: ChIP recovery in percentages of the input; right y-axis: fold induction for the RT-qPCR with reference to the untreated samples (t = 0 min)) and screen-shots from custom tracks of the UCSC genome browser for the ChIP-seq data (T0 and T30 only) for *CTGF* (A), *ZNF608* (B), *CDK5* (C), and *SERPINE1* (D). White bars: p300 ChIP; black bars: CBP ChIP; RT-qPCR data are indicated as dots, interconnected; arrows in the screen-shots indicate the position of the PCR-amplicon. Also indicated for these genes is the adjusted P-value and the ratio difference between time-points (T30 versus T0) and coactivators (p300 versus CBP) of the total number of reads along the whole gene, plus 1 kb up- or downstream from all ChIP-seq data (E).

show increased expression. The gene *ZNF608* shows a decrease in expression over time, consistent with decreased binding of p300/CBP. *CDK5* did not show any differences in expression. This is consistent with the uniform levels of CBP binding over time, but not with the increased binding of p300. Most likely, for *CDK5* and possibly also for other genes binding of p300/CBP is not sufficient to induce the expression but other factors that play a critical role are also required. Obviously, gene expression is a complex process and highly variable between genes, so only detailed studies can unravel the role of specific factors.

5.4.5 Biological Processes Coordinated by p300 and CBP

To get an impression of the biological implications of p300 and CBP binding, we clustered genes regulated by CBP and p300 into functional pathways. We used DAVID 2008 (89; 134) to classify the 250 genes most significantly differing between time points (for both CBP and p300). The analysis (p-value < 0.001) shows that CBP as well as p300 are mainly involved in transcription regulation of genes controlling developmental processes and metabolic processes (such as *NR4A*, *CRISPLD2*, *CRIM1*, *CYCLIN-L1*, and *PER1*) and of genes coding for proteins that control gene expression (such as *ATF3*, *FOSB*, *SP3* and *HES1*; see Additional File 5.8). Next, using DAVID 2008 we wanted to specify in more detail whether certain groups of genes were preferentially bound by CBP or by p300. Remarkably, in the cluster of genes regulating transcription, those with significantly higher CBP than p300 binding are involved in negative regulation of transcription (Table 5.4, and Additional File 5.8). Another interesting observation for genes preferentially bound by CBP is the presence of clusters related to signal transcription/cell communication. In the list obtained for higher levels of p300, mainly clusters relate to transcription and metabolic processes are found.

5.4.6 Analysis of ChIP-seq Regions for Consensus Transcription Factor Binding Sites

CBP and p300 do not bind DNA directly, but regulate by binding to many different protein partners. Therefore, to identify (DNA-binding) partners of p300/CBP, we looked for enrichment of TFBSs in and around the regions bound by CBP and/or p300 in the 250 genes that differ most significantly between time points (the same genes that were used for DAVID analysis). We found a significant over-representation of AP-1, CREB, NFKB and SRF binding sites in the gene regions bound by both CBP and p300 (Table 5.5 and Additional File 5.9), which are known to be regulated by CBP and/or p300 (136). As mentioned before, there is more binding of p300 and CBP to the chromatin at T30 after stimulation. Therefore, enrichment of the TFBSs in our sequences likely reflects increased binding of these factors upon growth factor stimulation.

We also compared genes significantly different between coactivators at T30 using the same lists of 250 genes as used for the functional classification. CREB and YY1 are significantly enriched in both gene sets (Table 5.5; all results are presented in Additional File 5.9). However, CBP binding was found to correlate more with AP-1 and SRF binding partners than p300, whereas p300 binding was more correlated to AP-2, E2F and SP1-binding. These results indicate that CBP and p300 share some, but not all, regulatory partners.

5.5 Discussion

Transcription coactivators CBP and p300 share high levels of homology and, in many cases, the same regulatory regions are targeted for transcription regulation. This is in contrast with the fact that both proteins are indispensable during embryogenesis. To investigate which genes are regulated, and whether there is a difference in those

Table 5.4: Functional classification for genes bound by CBP or p300

T30 p300 higher than CBP				Enrich-
ID (GO:#) Term	Count	%	p-value	ment
0010467 gene expression	81	38.76	2.57×10^{-12}	2.06
0044237 cellular metabolic process	127	60.77	3.75×10^{-10}	1.44
0008152 metabolic process	135	64.59	5.27×10^{-10}	1.38
0044238 primary metabolic process	126	60.29	1.36×10^{-9}	1.42
0043170 macromolecule metabolic process	111	53.11	8.79×10^{-8}	1.44
0006350 transcription	56	26.79	4.10×10^{-7}	1.95
0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	72	34.45	1.41×10^{-6}	1.66
0045449 regulation of transcription	53	25.36	1.82×10^{-6}	1.91
0016070 RNA metabolic process	58	27.75	3.07×10^{-6}	1.8
0019219 regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	53	25.36	3.64×10^{-6}	1.87
0010468 regulation of gene expression	54	25.84	4.93×10^{-6}	1.83
0031323 regulation of cellular metabolic process	54	25.84	1.62×10^{-5}	1.76
0043283 biopolymer metabolic process	84	40.19	1.78×10^{-5}	1.47
0019222 regulation of metabolic process	55	26.32	2.08×10^{-5}	1.73
0006351 transcription, DNA-dependent	48	22.97	3.29×10^{-5}	1.81
0032774 RNA biosynthetic process	48	22.97	3.39×10^{-5}	1.81
0006355 regulation of transcription, DNA-dependent	46	22.01	8.66×10^{-5}	1.77
0006979 response to oxidative stress	7	3.35	5.49×10^{-4}	6.83
0050794 regulation of cellular process	67	32.06	8.46×10^{-4}	1.42
T30 CBP higher than p300				Enrich-
ID (GO:#) Term	Count	%	p-value	ment
0051056 regulation of small GTPase mediated signal transduction	18	7.06	8.70×10^{-9}	5.97
0046578 regulation of Ras protein signal transduction	13	5.10	5.62×10^{-6}	5.36
0007242 intracellular signaling cascade	38	14.90	2.47×10^{-5}	2.06
0009966 regulation of signal transduction	21	8.24	2.49×10^{-5}	2.98
0007154 cell communication	77	30.20	3.31×10^{-5}	1.52
0007165 signal transduction	71	27.84	6.04×10^{-5}	1.54
0007265 Ras protein signal transduction	13	5.10	9.42×10^{-5}	4.03
0007264 small GTPase mediated signal transduction	18	7.06	1.32×10^{-4}	2.94
0007399 nervous system development	23	9.02	2.26×10^{-4}	2.4
0016481 negative regulation of transcription	13	5.10	3.31×10^{-4}	3.52
0045934 negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	13	5.10	7.31×10^{-4}	3.22

Significantly enriched GO categories for genes that show higher binding of CBP or p300 at 30 min after stimulation with TPA and serum (ID: GO-category-number, term: description of the GO category; count: number of significant genes in this GO category; %: percentage of signifiacnt genes in this GO category; p-value: statistical significance of the GO category (p-value from hypergeometric test for over-representation); enrichment: fold enrichment of significant genes compared to the background.

regulated by p300 and by CBP upon growth factor stimulation a genome-wide screen was performed in T98G cells. Although there is a high concordance between binding targets of p300 and CBP, and both seem to regulate the same biological pathways, we have identified significant differences in the levels and targets of binding. These differences include the diversity in the regulation of genes involved in transcription, and in cell death and cell adhesion. In addition, regulatory regions of these genes showed significant differences in binding sites of other TFs and TF families such as

Table 5.5: Enrichment for transcription factor binding sites in CBP/p300 bound sequences

T30 vs T0			T30		
TFBS	CBP	p300	TFBS	CBP>p300	p300>CBP
AP-1	0	0	AP-1	0	2.83×10^{-2}
CREB	8.21×10^{-5}	1.84×10^{-5}	AP-2	9.69×10^{-1}	7.30×10^{-9}
NFKB	4.55×10^{-7}	1.49×10^{-4}	CREB	2.96×10^{-4}	7.60×10^{-9}
SRF	3.41×10^{-7}	1.51×10^{-5}	E2F	2.39×10^{-1}	0
			SP1	9.98×10^{-1}	2.21×10^{-7}
			SRF	8.81×10^{-4}	2.12×10^{-1}
			YY1	2.55×10^{-6}	0

TFBSs with the most significant p-values for enrichment in regions bound by CBP and p300 at 0 and 30 min after stimulation with serum and TPA.

AP-1, AP-2, SP1, E2F and SRF.

It is well established that p300/CBP associate to both enhancers and TSSs. Previous studies have focused on the enhancer-binding of p300 (119; 118; 117). Although we also found examples of enhancer-binding, over 57% of all tags are within genes or proximal promoters ($+/-1$ kb), and genome-wide we find that binding is primarily located around TSSs and to some extend also to transcript ends (Figure 5.1A and B). Therefore, we chose to focus our analysis of CBP/p300 in relation to genic regions. In all, we found 16,103 genes bound by CBP or p300 at T0 or T30, with over 97.4% of genes bound by both coactivators at both time points.

When analyzing the binding of CBP/p300 to genes, we did not only observe distinct regions of binding. There was a high variety in binding patterns for both coactivators. This includes binding to a clear and distinct region (*e.g.* to the TSS; referred to as peak), binding across the gene, or a combination of more prominent binding around the TSS and the transcript end, as well as binding across the gene (in the text referred to as U-shaped binding; Figure 5.1C-E).

At present, the mechanisms that determine the diverse binding patterns remain to be established. Possibly, it is dependent on the way p300/CBP regulate transcription of a particular gene. Both, p300 and CBP can bind to specific TFs, and this might result in a distinguished peak around the TSS and transcript end. In addition, p300 and CBP regulate chromatin structure via the acetylation of histones, thereby making the chromatin more prone to be targeted for transcription. This might account for binding (to the histones) across the gene. Binding across a gene was previously described to occur also by protein kinases (137). Chow *et al.* (137) propose that in this way the kinases may contribute to transcription initiation and elongation, or processes such as $5'$ capping, and splicing. Binding to both the TSS and transcript end has previously been observed for RNA polymerase II (138). Interaction between CBP/p300 and RNA polymerase II (139) may explain the presence of similar ChIP-seq patterns for these acetyltransferases. Enrichment at the TSS might correlate to the longer time needed for transcription initiation compared to transcript elongation. The peak at the transcript end might correlate to widespread transcription of antisense transcripts (140), a phenomena that is particularly prominent in the $3'$-end of genes (36).

Our ChIP-seq data are from arrested cells and from cells 30 min after stimulation. Therefore, over-representation of genes required early in the cell cycle was expected at T30. The number of reads correlates roughly to the binding affinity of proteins for that region and immediate-early genes are among the genes with the highest number of tags. Analysis of a number of these genes with quantitative RT-PCR (Figure 5.3A, Additional Figure 5.5, and data not shown), also showed increase in gene expression for immediate-early genes. Our data suggest that at T30 CBP and p300 are more intimately involved in the regulation of transcriptional activation of immediate-early genes compared to other groups of genes. Consistently, Tullai *et al.* (131) previously published microarray data on gene expression of serum-starved T98G cells upon growth stimulation. We found that from 49 immediate-early genes that were identified, 36 demonstrated significantly increased binding by CBP and p300 at T30 compared to T0 in our analysis (3 out of 49 could not be identified in Ensembl).

With the time-course experiment, most genes that were analyzed show maximal binding between 30 and 60 min after stimulation. The time-course experiment confirms high accuracy of our data since all genes tested, although different levels of significance (from 2×10^{-147} to 9×10^{-1}) and variable ratios of difference (from 1 to 9) were chosen, confirm binding of the coactivators and changes in time.

Binding of p300 and CBP to the chromatin occurs through the interaction with TFs. To obtain more insight in transcription regulatory complexes bound by p300 and/or CBP, we set out to identify possible partners of CBP and p300 for the genes identified in our experiment. Therefore, we analyzed for the enrichment of TFBSs. When looking at genes with significant binding at T30 for each coactivator, we found some examples of TFBSs that were found to be specific only for CBP or for p300. For example, AP-1 and SRF binding sites were significantly enriched in CBP bound regions, while AP-2, E2F and SP1 binding sites were more abundant in p300 bound regions. This may represent TFs that are regulated during the cell cycle, in most cases, solely by CBP or p300 and contribute to their unique functions.

We observed overlap in enrichment of TFBSs for proteins such as YY1 and CREB. Interestingly, YY1 is known to contribute to cell-cycle regulation and can serve both, as a transcriptional repressor and an activator (141). Also, YY1 is known to interact with p300/CBP, as well as with other TFs identified in this study (AP-1, AP-2, NFKB, E2, SP1 and CREB) (141; 142; 143). Our functional classification suggested that CBP is more associated with transcriptional repression, whereas p300 is more associated with transcriptional activation. It could be speculated that YY1 is a putative partner involved in this functional difference between p300 and CBP, while the p300-YY1 complex might activate transcription *in vivo*, the CBP-YY1 complex might account for transcriptional repression.

In the future, it would be valuable to perform ChIP-seq in the same cell line and conditions with antibodies for the coactivator specific TFs in this study (AP-1, AP-2, SP1, E2F and SRF, and YY1). This will confirm whether genome-wide CBP/p300 and their specific regulatory partners cooperate, and it will help to further elucidate their role in cell-cycle control. In addition, ChIP-seq with antibodies specific to open chromatin states will be helpful to unravel the mechanisms leading to the diverse binding patterns.

5.6 Conclusion

Transcription coactivators CBP and p300 share high levels of homology and, in many cases, the same regulatory regions are targeted for transcription regulation. This is in contrast with the fact that both proteins are indispensable during embryogenesis. To investigate which genes are regulated, and whether there is a difference in those regulated by p300 and by CBP upon growth factor stimulation a genome-wide screen was performed in T98G cells. Although there is a high concordance between binding targets of p300 and CBP and both seem to regulate the same biological pathways, we have identified significant differences in the levels and targets of binding. In addition, regulatory regions of target genes also showed significant differences in TFBSs of other TFs such as AP-1, AP-2, SP1, E2F and SRF.

Besides the differences in targets of p300 and CBP, we identified various binding-patterns that potentially correlate with different types of transcription regulation by p300 and CBP. Most interestingly, we observed a so-called U-shaped binding with high levels of p300/CBP at both, TSS and transcript end. Possibly, the acetyltransferases contribute to other processes such as transcription elongation and reverse transcription. Taken together, our data contribute to the improvement of our knowledge of processes that regulate gene expression by the transcription coactivators p300 and CBP, and confirm that regulation by these coactivators is not identical.

5.7 Funding

Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO); Center for Biomedical Genetics (in the Netherlands).

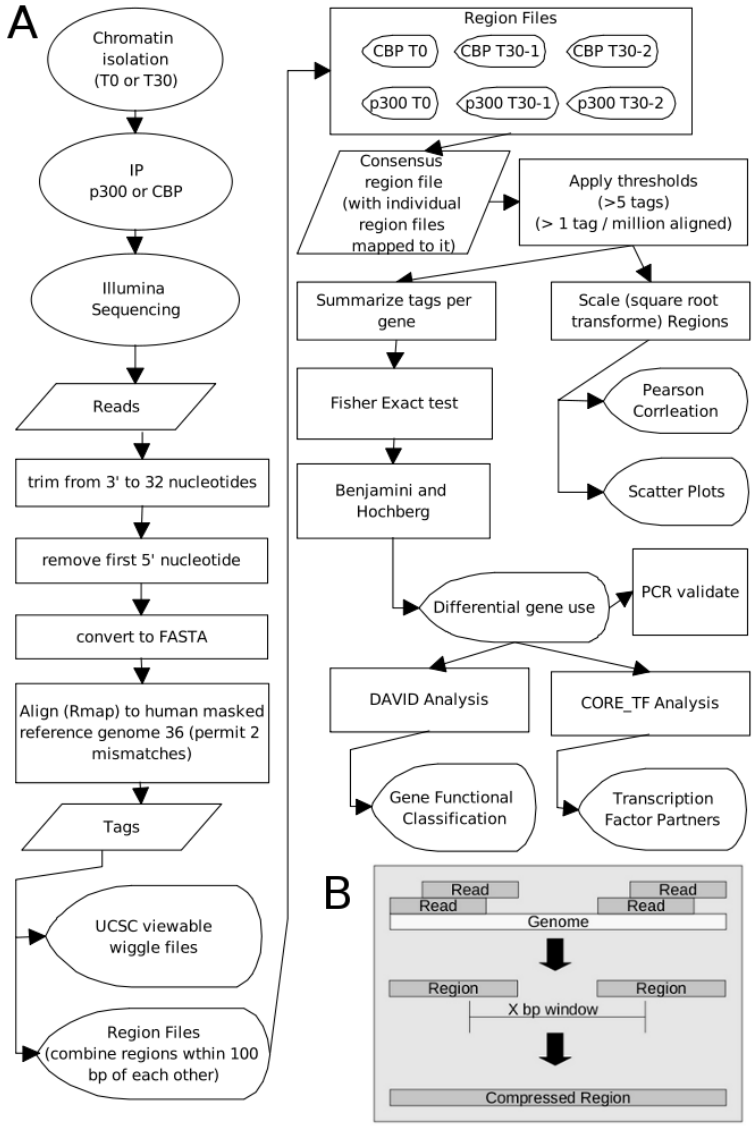
Conflict of interest statement. None declared.

5.8 Acknowledgements

We wish to thank Michel P. Villerius for his computational support and Dorien JM Peters for her review and comments about the manuscript.

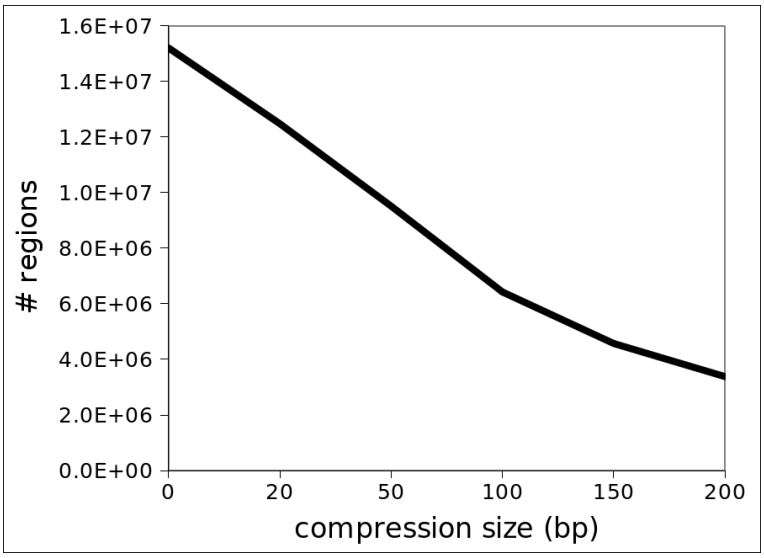
5.9 Additional Files

Additional Figure 5.1



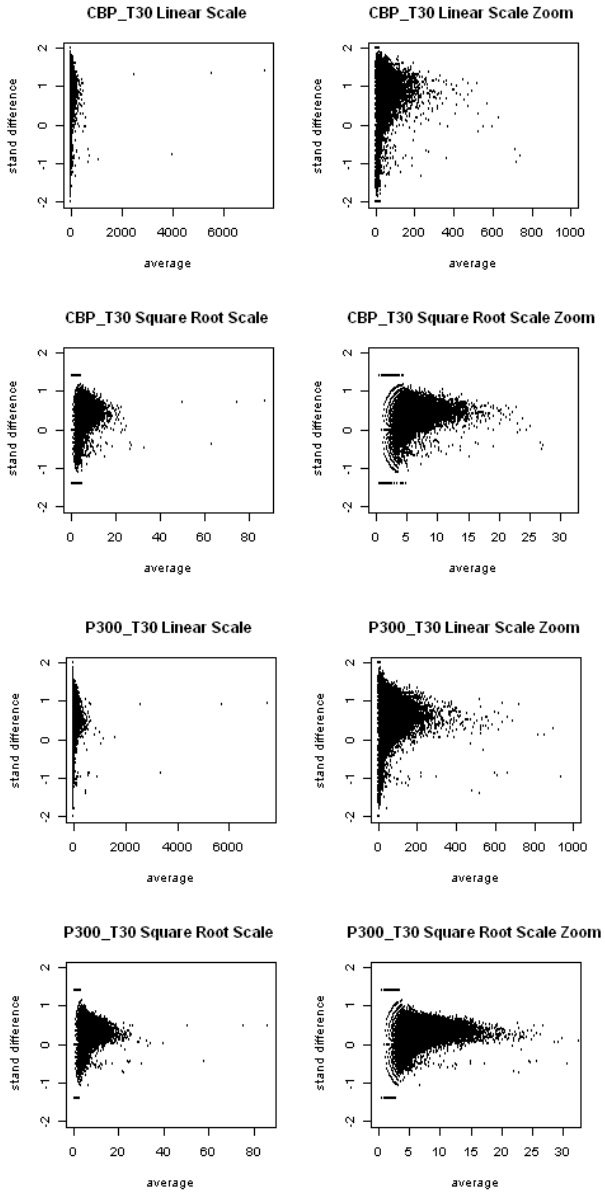
Flow chart of the experimental set-up (A). Also demonstrated is the creation of regions from aligned reads and compressing regions within a window of X bps (B).

Additional Figure 5.2



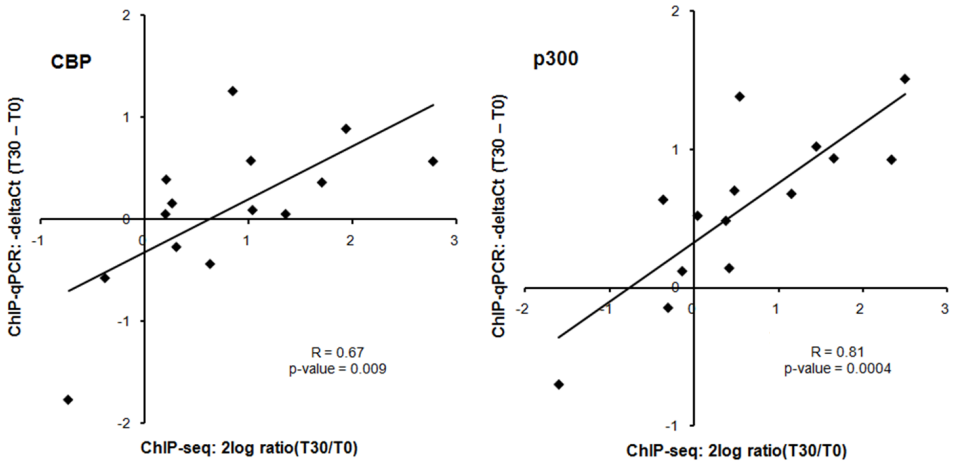
To demonstrate the effect of compressing regions with variable window sizes (x-axis) we plotted the number of regions obtained (y-axis).

Additional Figure 5.3



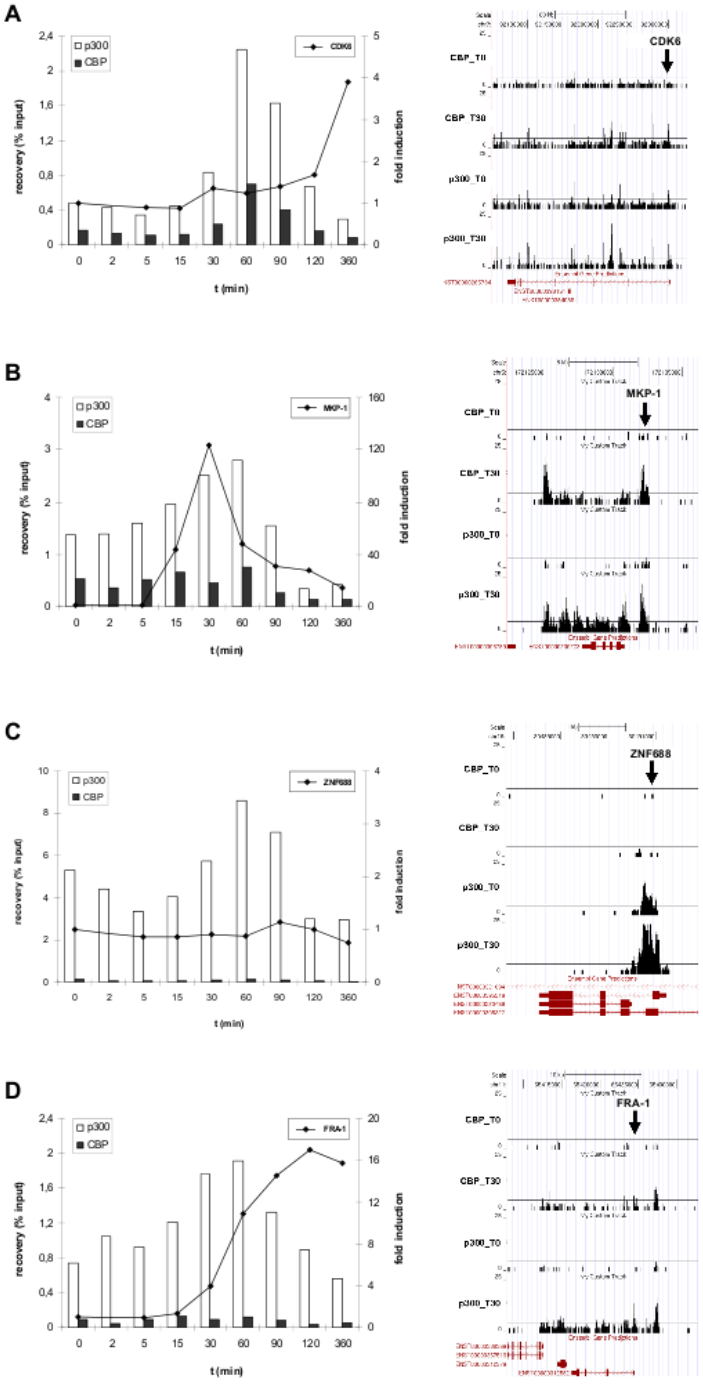
To demonstrate the variance stabilizing property of the square root transformation, we plotted the standardized difference (=difference divided by the mean) of the tags per region in the two biological replicates for CBP_T30 (A) and P300_T30 (B) against the average number of tags per region for those samples. Top panels are on the linear scale. Lower panels are on the square root scale. Left panels show the entire range of tags; right panels zoom in on the majority of regions with lower number of tags. The plot shows that the variance is much more homogeneously distributed on the square root scale.

Additional Figure 5.4



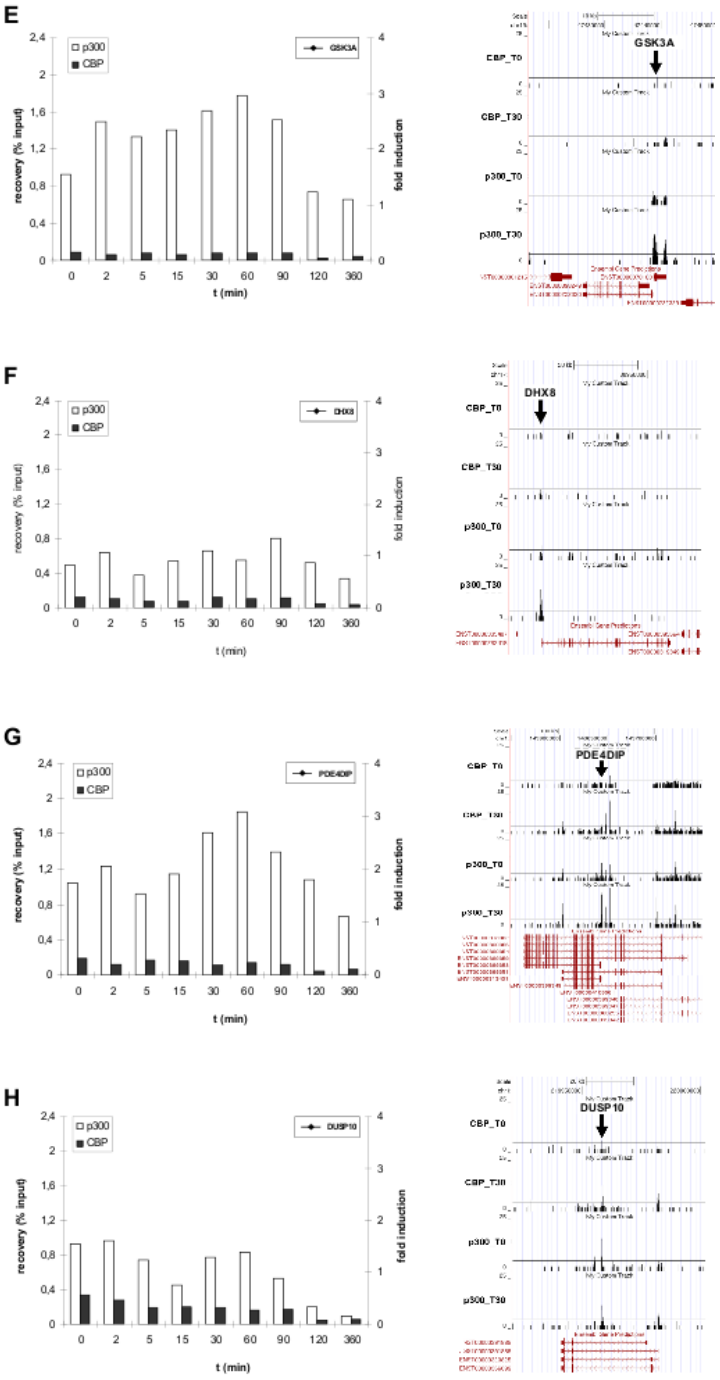
Correlation between ChIP-seq and ChIP-qPCR data for CBP (left) and p300 (right). For all genomic regions validated by qPCR (*CDK5*, *CDK6*, *CTGF*, *DHX8*, *DUSP10*, *FRA1*, *GSK3A*, *MKP1*, *PDE4DIP*, *SERPINE1* (TSS), *SERPINE1* (2kb upstream), *TAF15*, *ZNF608*, and *ZNF688*), we plotted the T30/T0 ratio obtained from the ChIP-Seq experiments (x-axis) against the T30/T0 ratio obtained from the ChIP-qPCR experiments (y-axis). Since deltaCt values plotted for the qPCR experiments reflect 2log differences in binding, also the ratio of the number of sequences from the ChIP-Seq experiments were 2log transformed. We counted the number of sequences aligned to each bp in the region spanned by the PCR primers +/- the average fragment length of 500 nucleotides. This was done because only the starts of fragments were sequenced and aligned, and these may fall outside of the PCR region, despite the presence of the PCR region in the fragment. We plotted delta Ct (y-axis) to provide a positive correlation since lower Ct values represent more binding, and higher Ct values less binding. The Ct values for time point T0 and T30 were obtained after fitting of a second-order polynomial through the Ct values of the time course from 0 to 90 minutes to improve the precision. The Pearson correlation coefficient and the p-value representing the significance of the correlation are given.

Additional Figure 5.5A-D



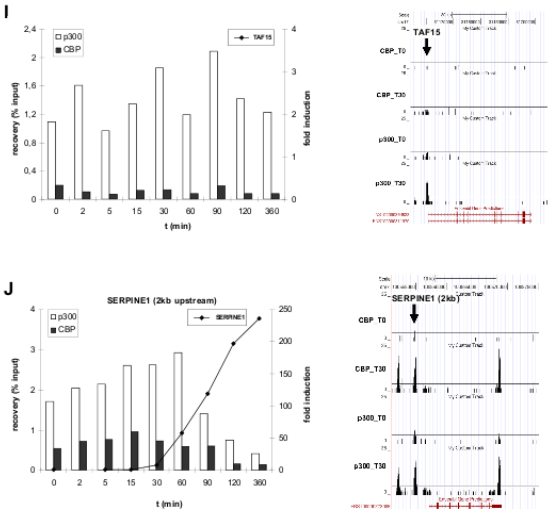
continued on next page

Additional Figure 5.5E-H



continued on next page

Additional Figure 5.5I-K



K

Gene	CBP_T30vsT0		p300_T30vsT0		T0_p300vsCBP		T30_p300vsCBP	
	adj.p-value	ratio	adj.p-value	ratio	adj.p-value	ratio	adj.p-value	ratio
CDK6	3.60 x 10^{-2}	0.92	3.48 x 10^{-26}	0.78	3.48 x 10^{-2}	1.10	2.42 x 10^{-3}	0.93
MKP-1	2.28 x 10^{-41}	3.94	2.05 x 10^{-147}	6.81	9.80 x 10^{-1}	0.95	1.47 x 10^{-35}	1.65
ZNF688	9.41 x 10^{-7}	1.02	7.25 x 10^{-13}	1.42	3.84 x 10^{-2}	1.26	9.84 x 10^{-2}	1.75
FRA-1	1.96 x 10^{-3}	3.36	8.86 x 10^{-34}	5.60	1.00	1.01	2.08 x 10^{-3}	1.68
GSK3A	8.39 x 10^{-1}	0.86	1.49 x 10^{-4}	1.71	5.38 x 10^{-2}	2.00	1.16 x 10^{-25}	3.96
DHX8	5.70 x 10^{-3}	1.71	1.85 x 10^{-3}	0.59	4.55 x 10^{-1}	1.21	7.07 x 10^{-31}	0.42
PDE4DIP	2.59 x 10^{-1}	0.94	1.95 x 10^{-4}	0.89	1.00	0.99	2.66 x 10^{-2}	0.94
DUSP10	9.03 x 10^{-1}	0.97	7.84 x 10^{-2}	0.87	2.24 x 10^{-1}	1.21	2.74 x 10^{-1}	1.09
TAF15	6.77 x 10^{-2}	0.73	6.74 x 10^{-7}	1.07	8.07 x 10^{-1}	1.11	1.13 x 10^{-7}	1.62
SERPINE1	1.88 x 10^{-17}	3.60	1.59 x 10^{-24}	3.57	1.00	1.03	9.29 x 10^{-7}	1.02

(continued from previous page): ChIP-analysis for time course experiment (0, 2, 5, 15, 30, 60, 90, 120, and 360 minutes after stimulation of growth arrested T98G cells with serum and TPA). Shown are graphs for qPCR results (x-axis: time in minutes; left y-axis: ChIP recovery in percentages of the input; right y-axis: fold induction for the RT-qPCR with reference to the untreated samples (t=0 minutes)) and screen-shots from custom tracks of the UCSC genome browser for the ChIP-seq results (T0 and T30 only) for *CDK6* (A), *MKP-1* (B), *ZNF688* (C), *FRA-1* (D), *GSK3A* (E), *DHX8* (F), *PDE4DIP* (G), *DUSP10* (H), *TAF15* (I), and *SERPINE1* (J; 2 kb upstream of the TSS). Arrows indicate the position of the PCR-amplicon. Also indicated for these genes is the adj. p-value and the ratio difference between time points (T30 versus T0) and coactivators (p300 versus CBP) of the total number of sequences from all ChIP-seq data (K). (White bars: p300 ChIP; black bars: CBP ChIP; line: fold induction of RT-qPCR; arrows in the screen-shots indicate the position of the PCR-amplicon. Also indicated for these genes is the adjusted p-value and ratio difference between time-points (T30 versus T0) and coactivators (p300 versus CBP) of the total number of reads from all ChIP-seq data (K)).

Additional Table 5.1

A

primer:	sequence:
ChIP CDK5 Fw	5'-TGCATTCTGGAACGCGTAGTC-3'
ChIP CDK5 Rv	5'-TTATGTGCTCCCAGCCCTCTT-3'
ChIP CDK6 Fw	5'-CCATGCTTTTCCCAGATGGA-3'
ChIP CDK6 Rv	5'-GAGTAAGAACTGCTGTCTCCCCA-3'
ChIP CTGF Fw	5'-TGAGTTGATGAGGCAGGAAGG-3'
ChIP CTGF Rv	5'-CACAAACAGGGACATTCCTCG-3'
ChIP DHX8 Fw	5'-GCGGAAGAACTTGCCAAACTC-3'
ChIP DHX8 Rv	5'-CCCAAGTGATTGTCCAGCTCA-3'
ChIP DUSP10 Fw	5'-TCCCCCTCCTTTAATCCGTCT-3'
ChIP DUSP10 Rv	5'-GGCTCCACCTTTTTTGCTCA-3'
ChIP FRA-1 Fw	5'-AAGGCCAGTGGAAAGACCTCA-3'
ChIP FRA-1 Rv	5'-CCGTTTCTGCTCCCACAAAA-3'
ChIP GSK3A Fw	5'-TCAGGTCTGGACTATTCCCCA-3'
ChIP GSK3A Rv	5'-GTTCGACGTATCCTCCCCATT-3'
ChIP MKP-1 FW	5'-CTTTGCTGTCCTCGACCAATG-3'
ChIP MKP-1 RV	5'-CTGCGCAAATGTTCTGCCT-3'
ChIP PDE4DIP Fw	5'-CTTAGCAGATGAAAGCGGCTG-3'
ChIP PDE4DIP Rv	5'-AAATCCGGTTGCACACCTG-3'
ChIP SERPINE1 Fw	5'-AAAAGCAGGCAACGTGAGCT-3'
ChIP SERPINE1 Rv	5'-TGACCCAAAAAGCCTAGGACC-3'
ChIP SERPINE 1 (2kb) Fw	5'-TCAGTGTGGTTGCCCTTGGTA-3'
ChIP SERPINE (2kb) 1 Rv	5'-GGAAGCGTCGGATGTTTGTT-3'
ChIP TAF15 Fw	5'-CCTCTTTCGTTTCCTAACCGC-3'
ChIP TAF15 Rv	5'-AGTACCAATGCCACGATCACG-3'
ChIP ZNF608 Fw	5'-TGACTAGCACAGCCGCACTTT-3'
ChIP ZNF608 Rv	5'-TGTGCCTATTTTCCCCTTCG-3'
ChIP ZNF688 Fw	5'-TTGCGTATTGCGGGTACAGAG-3'
ChIP ZNF688 Rv	5'-GAAAGATGTAAGGGCCCGAAG-3'

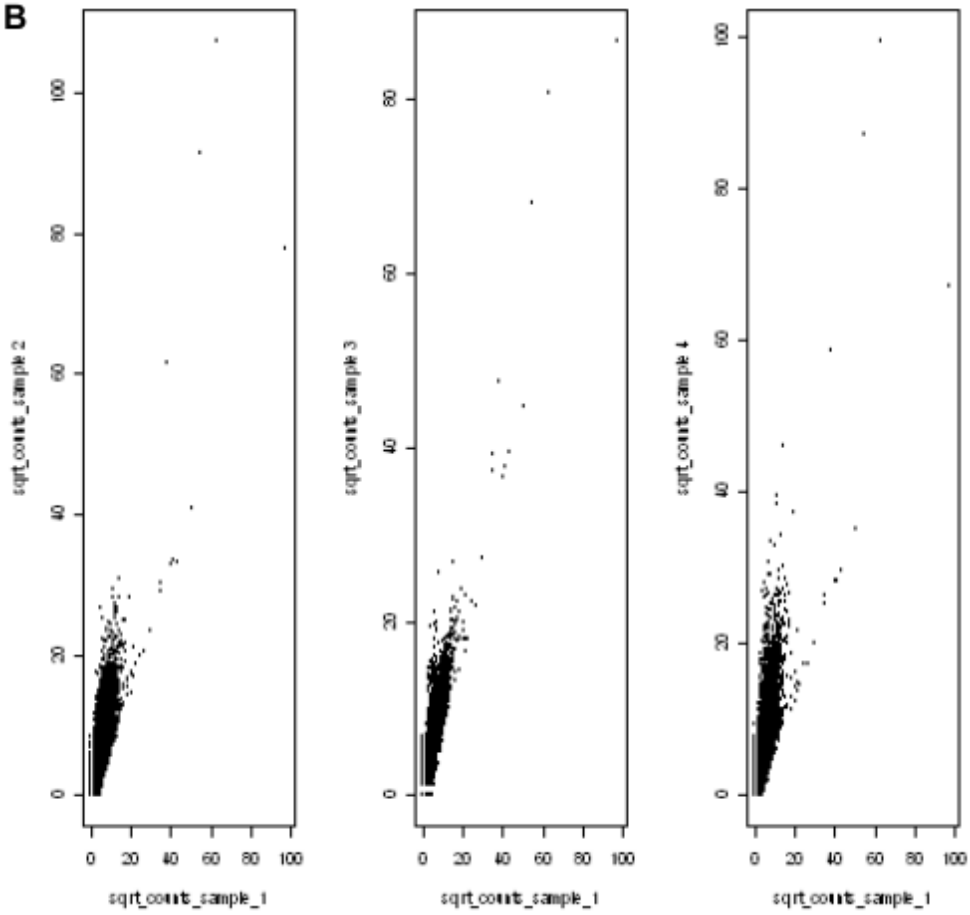
B

primer:	sequence:
RT CDK5 up	5'-GCGACAAGAAGCTGACTTTGG-3'
RT CDK5 dn	5'-CGAGGTCACCATTGCAACTGT-3'
RT CDK6 up	5'-CTGGCATTCAATCGTTGGC-3'
RT CDK6 dn	5'-TACACCGATGGAAGAACTGGG-3'
RT CTGF up	5'-TACCAATGACAACGCCTCCTG-3'
RT CTGF dn	5'-ACGGATGCACTTTTTGCCC-3'
RT FRA-1 up	5'-CTCTTCCTGTGATCCACCCAA-3'
RT FRA-1 dn	5'-TGGGTAAAGTGGCACCTTCTG-3'
RT MKP-1 up	5'-CTTTGCTGTCCTCGACCAATG-3'
RT MKP-1 dn	5'-CTGCGCAAATGTTCTGCCT-3'
RT SERPINE1 up	5'-TGGCCCATGAAAAGGACTGT-3'
RT SERPINE1 dn	5'-AGAGAACCTGGGAATGACCGA-3'
RT ZNF608 up	5'-CGTGGAGGAACAAAACGTACG-3'
RT ZNF608 dn	5'-GGTGACTCACAAAACCTGGGA-3'
RT ZNF688 up	5'-TTGCGTATTGCGGGTACAGAG-3'
RT ZNF688 dn	5'-AAGAGATGAGGGCTGGTTTGG-3'

Sequence of primers used for qPCR analysis of the ChIPs (A) and for RT-qPCR (B).

Additional Table 5.2

A) Pearson Correlation	CBP T0	CBP T30[1]	CBP T30[2]	p300 T0	p300 T30[1]	p300 T30[2]
CBP T0	1	0.6698245	0.7215659	0.7663312	0.660192	0.722461
CBP T30[1]		1	0.7654689	0.740022	0.869117	0.85208
CBP T30[2]			1	0.7202088	0.773906	0.805657
p300 T0				1	0.753338	0.808835
p300 T30[1]					1	0.869657
p300 T30[2]						1



Pearson correlation between the different samples (A) and scatter plots (B).

Additional Table 5.3

Fisher Test	adj.p-value <0.01	adj.p-value <0.001	difference >= 5x
CBP T30vsT0	1231	765	11
p300 T30vsT0	3730	2620	44
T0 p300 vs CBP	256	120	22
T30 p300 vs CBP	2502	1611	42

Number of genes significantly different between the different samples for adjuvant p-values <0.01 and <0.001, as determined with Fisher Exact Test, and the number of genes with p-values <0.001, with a difference of at least 5 times between the samples.

Additional Files 5.1 to 5.6
Wiggle files for CBP T0, CBP $T30^1$, CBP $T30^2$, p300 T0, p300 $T30^1$, and p300 $T30^2$, created as described in Materials and Methods and viewable in the UCSC genome browser ((103) (single reads were removed) are available at *nar.oxfordjournals.org/cgi/content/full/gkq184/DC1.*

Additional File 5.7
Genes annotated for p300 and CBP ChIP-seq in quiescent (T0) and in growth factor stimulated (T30) cells. Sheet 1 shows the 1,6103 genes that were identified (as explained in Sheet 2, the Ensemble gene ID, the sum of the number of tags sequenced and the number after normalization for gene length, the ratio between the different samples, the p-value, and the adjusted p-value for the ratios are shown). Available at *http://nar.oxfordjournals.org/content/vol0/issue2010/images/data/gkq184/DC1/NAR-02256-X-2009_R2_supplemental_file_7.xls.*

Additional File 5.8
Functional classification performed with DAVID 2008 Functional Annotation (http://david.abcc.ncifcrf.gov/home.jsp) of 250 genes that differed most significantly at T30 versus T0 for binding by CBP (worksheet CBP_T30vsT0), by p300 (worksheet p300_T30vsT0), and of 250 genes where CBP binding is significantly higher than p300 binding (worksheet T30_CBPhigherthanp300) or where p300 binding is significantly higher than CBP binding (worksheet T30_p300higherthanCBP). Available at *http://nar.oxfordjournals.org/content/vol0/issue2010/images/data/gkq184/DC1/NAR-02256-X-2009_R2_supplemental_file_8.xls.*

Additional File 5.9
Full list of Transcription Factor binding sites that were found to be enriched for T30 versus T0, and for p300 versus CBP, as analyzed with CORE_TF (76). Available at *http://nar.oxfordjournals.org/content/vol0/issue2010/images/data/gkq184/DC1/NAR-02256-X-2009_R2_supplemental_file_9.xls.*

Chapter 6

Tissue Specific Transcript Annotation and Expression Profiling with Complimentary Next-generation Sequencing Technologies

Matthew S. Hestand [1,2], Andreas Klingenhoff [3], Matthias Scherf [3], Yavuz Ariyurek [2], Yolande Ramos [4], Wilbert van Workum [5], Makoto Suzuki [6], Thomas Werner [3], Gert-Jan B. van Ommen [1], Johan T. den Dunnen [1,2], Matthias Harbers [6], Peter A.C. 't Hoen [1]

[1]The Center for Human and Clinical Genetics, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[2]Leiden Genome Technology Center, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
[3]Genomatix Software GmbH, Munich, Germany.
[4]Department of Molecular Cell Biology, Leiden University Medical Centre, 2300 RC Leiden, The Netherlands.
[5]ServiceXS B.V., Plesmanlaan 1D, 2333 BZ Leiden, The Netherlands
[6]DNAFORM Inc., Leading Venture Plaza-2, 75-1, Ono-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0046, Japan.

manuscript submitted

6.1 Abstract

Next-generation sequencing is excellently suited to evaluate the abundance of mRNAs to study gene expression. Here we compare two alternative technologies, cap analysis of gene expression (CAGE) and serial analysis of gene expression (SAGE), for the same RNA samples. Along with quantifying gene expression levels, CAGE can be used to identify tissue-specific transcription start sites, while SAGE monitors 3′ end usage. We used both methods to get more insight into the transcriptional control of myogenesis studying differential gene expression in differentiated and proliferating C2C12 myoblast cells with statistical evaluation of reproducibility and differential gene expression. Both CAGE and SAGE provided highly reproducible data (Pearson correlations > 0.92 between biological triplicates). With both methods we found around 10,000 genes expressed at levels > 2 transcripts per million ($\sim$0.3 copies per cell), with an overlap of 86%. We identified 4,304 and 3,846 genes differentially expressed between proliferating and differentiated C2C12 cells by CAGE and SAGE respectively, with an overlap of 2,144. We identified 196 novel regulatory regions with preferential use in proliferating or differentiated cells. Next-generation sequencing of CAGE and SAGE libraries provides consistent expression levels and can enrich current genome annotations with tissue-specific promoters and alternative 3′ UTR usage.

6.2 Introduction

Next-generation sequencing (NGS) platforms have provided us with the technology needed to expand genomic methods to a new scale. Depending on the technology, these machines can produce gigabases of sequences per day. Due to its superior resolution and sensitivity, NGS is increasingly used to replace array technologies, in particular the genome-wide evaluation of chromatin immunoprecipitation (ChIP-seq) and gene expression profiling experiments. Sequence-based expression analysis can be performed using several approaches. The traditional SAGE (serial analysis of gene expression) method (28), starts with capturing RNA poly-A tails with oligo(dT) beads. Double-stranded cDNA synthesis is performed and a digestion with a restriction enzyme, commonly NlaIII (32), is performed. With the fragments resulting from the digestion only the most 3′ fragment is retained. An additional restriction digest is then performed with MmeI (cuts ∼20 base pairs downstream) to create a fragment of acceptable length for sequencing. In the original method short cDNA fragments, each representing the 3′ most NlaIII digestion site of a specific transcript, were concatenated and cloned, followed by traditional sequencing. However, now the concatenation and cloning steps can be omitted. Instead SAGE library sequences are directly equipped with appropriate sequencing linkers and analyzed in next-generation sequencers (30).

An alternative method is CAGE (cap analysis of gene expression) (29), specifically designed to study gene expression at transcription initiation sites by capturing 5′ ends of mRNAs. After trapping the 5′ cap-structures of mRNAs, sequences are converted to double-stranded cDNA and equipped with a linker containing a restriction site for the enzyme MmeI (or EcoP15I) that cuts ∼20 (or 25-27) base pairs downstream to create a fragment of appropriate length for sequencing and for mapping. Thus where SAGE captures the 3′ most NlaIII digestion site of mRNA and is thus 3′ end biased, CAGE tags represent the ultimate 5′ end of the transcript and indicate the genomic transcription start site (TSS). In both SAGE and CAGE, one transcript is only represented by a single read and (next-generation) sequencing of SAGE and CAGE libraries is therefore referred to as Digital Gene Expression profiling or DeepSAGE and DeepCAGE (30; 31). For simplicity we refer to these in this manuscript simply as SAGE and CAGE. In RNASeq (144), which starts with random fragmentation of the RNA or cDNA, the entire transcript is sequenced. Consequently, a transcript is commonly represented by multiple reads and the amount of reads is dependent on the transcript length. RNASeq gives more detailed information about the structure of the transcripts and alternative splicing, in particular when combined with paired end sequencing, while CAGE is more suitable for analysis of alternative transcription start sites and SAGE for analysis of alterantive polyadenylation sites.

Myogenesis is an essential process for muscle development and regeneration, with defects resulting in diseases such as muscular dystrophies. To support our studies towards treatment of muscle-related diseases, we have performed extensive analysis of muscle-derived gene expression profiles (145; 146; 147). This included the analysis of muscle differentiation using a well-established model, the mouse myoblast cell line (C2C12) (148). Two primary transcription factors (TFs) regulating this process are MyoD and Myogenin, but many other regulatory elements have been identified (reviewed in Pownall *et al.* 2002 (24) and Sartolli and Caretti 2005 (149)). For a better

understanding of how expression profiles change during adaptation to different biological situations, it is important to consider promoter activities and their regulation. Several bioinformatic approaches have been designed for this, including CORE_TF (76) and oPOSSUM (60), searching for shared TF binding sites (TFBSs) in the promoter region. However, these approaches critically depend on correct genome annotations regarding TSSs, which can vary by tissue type. Unfortunately, most studies performed thus far use methods directed at the 3′ end of RNA transcripts (including the well known oligo dT primed cDNA synthesis). Consequently gene annotation is weakest at the 5′ end. CAGE is therefore excellently suitable for the identification of alternative TSSs and putative regulatory regions upstream of those TSSs. We applied both CAGE and SAGE to study muscle differentiation to assess their concordance in estimation of gene expression levels and complementarity in gene annotation.

6.3 Materials and Methods:

6.3.1 Cells, RNA Isolation, and Differentiation Markers

Proliferating C2C12 mouse myoblasts were grown out on collagen coated plates in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS). To induce fusion into myotubes cells were serum deprived by changing to a medium of DMEM supplemented with 2% FBS for nine days (referred to as differentiated cells).

For CAGE and SAGE, RNA was isolated from proliferating and differentiated cells. RNA was isolated from three independent cultures (biological triplicates). Cells grown in (175 cm^2) flasks where harvested by trypsinization and centrifugation before RNA extraction with a Nucleospin RNA L kit from Macherey-Nagel. RNA quality was high, as determined with Agilent's Lab-on-chip total RNA nano assay (RNA integrity number >9). Myogenic properties of the cells were confirmed in RT-PCR/qPCR experiments using primer sets (Additional Table 6.1) specific for *Myod1*, *Myogenin*,*GAPDH*, and *HPRT*. RT-PCR experiments were performed using oligo dT priming for cDNA synthesis and qPCR was carried out using a Roche Lightcycler 480.

6.3.2 Library Preparation and Next-Generation Sequencing

Separate CAGE libraries were prepared as described previously (31) for each individual RNA sample. The following modifications to the protocol were made: we used modified adapters in the 5′ and 3′ end ligation steps that have linker sequences (proliferating - CCGACAGGTTCAGAGTTCTACAGAGACAGCAG and differentiated - CCGACAGGTTCAGAGTTCTACAGCTTCAGCAG) for Illumina Genome Analyzer II sequencing and have a recognition site for EcoP15I used instead of MmeI.

SAGE libraries were prepared for each individual RNA sample with a FC-102-1005 DGE-Tag Profiling NlaIII SamplePrepKit from Illumina.

Each CAGE and SAGE library was then sequenced on an individual lane on an Illumina Genome Analyzer II for 36 cycles. One CAGE sample from each time point was also sequenced a second time with 32 cycles.

6.3.3 Initial Sequence Analysis

All sequenced lanes were run through the initial Illumina Genome Analyzer Pipeline (Firecrest ⇒ Bustard ⇒ Gerald) for image analysis and quality control, yielding one scarf file per sample (lane). For reads from SAGE samples, the NlaIII recognition sequence "CATG" was introduced at the 5′-end with Linux commands. Scarf files were then run through the open source GAPSS_R pipeline developed in house (*www.lgtc.nl/GAPSS*). In general, this pipeline takes sequences and has the options to: remove first bases (often of lower quality than other 5′ nucleotides (102)), edits for linkers (present in the sequence reads when sequencing more cycles than the fragment length), aligns to a reference genome with Rmap (40), and reports data as region files (reporting tags in a region, a region defined as a stretch of adjacent nucleotides with aligned reads), and creating UCSC genome browser (103) (*http://genome.ucsc.edu/*) viewable wiggle tracks.

We ran GAPSS_R with the parameters discussed in the following text. The first base (lower quality) was removed in CAGE samples. CAGE and SAGE samples were edited for 3′ linker sequences (TCGTATGCCGTCTTCTGCTTG for CAGE and TCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAA for SAGE), permitting 1 mismatch in the linker (to account for sequencing errors, which occur more towards the 3′ end (102) where linkers were edited from). After linker editing, the majority of CAGE reads were 26 bases in length, whereas SAGE reads were 21 or 22 bases in length (including the "CATG"). Alignment was performed against the mouse repeat masked reference genome build 37 with Rmap v0.41, an alignment tool that reports only unique alignments. Default settings were used during alignment, except to use fasta input and permitting 2 mismatches with CAGE reads and 1 mismatch with SAGE reads. The choice of mismatches permitted is because longer sequences (CAGE) are more likely to contain a sequencing error because the number of errors increases at later sequencing cycles. Region files were created and for CAGE regions we combined adjacent regions, permitting gaps of maximal 100 bases to cluster TSSs and make sure that newly identified TSSs were well separated from annotated TSSs. We kept all data separated by strand since both methods preserve information on the transcribed strand. Wiggle files for visualization in the UCSC genome browser were also separated by strand.

Custom Perl scripts were run on all CAGE and SAGE region files to create reference region files (strand separated) composed of the overlapping regions from all samples. For CAGE region files we again permitted gaps of a maximum by 100 bases. Another custom Perl script was used to link all individual region files to their reference region file, reporting the estimated number of tags in each individual region of the reference region file.

6.3.4 Statistical and Biological Processes Analysis

The statistical language R was then used for analysis of differential expression for CAGE and SAGE data. A threshold of two tags per million aligned reads (average across all samples) was applied to remove transcription events that do not pass the lower limit for consistent detection given our read depth. In addition, for CAGE data, we excluded regions of length 33 or lower. These are likely sample preparation artifcats, since these were usually caused by exactly identical reads of 33 nt, which did

not contain the linker sequence. Even sharply defined TSSs demonstrate variability in start position, resulting in regions that cover >33 nt. Each region was tested separately with a Bayesian algorithm that takes into account library size (150; 36). A Bayesian error rate lower than 0.05 was considered significant. For gene level tests, all tags overlapping a gene (including 1000 bases upstream and downstream of the gene) were summarized before statistical testing. For the calculation of expression ratios between differentiated and proliferating cells, data was first scaled to the average total number of aligned reads. For analysis of reproducibility, data was square root transformed to stabilize variance between samples, after which the Pearson correlation coefficient was calculated.

To compare differentially expressed genes to previously published microarray data we took results from Tomczak *et al.* 2004 (148), performed VSN normalization (151), and analyzed data from differentiated versus proliferating cells with limma (67; 68) in R. Multiple testing was done according to Benjamini and Hochberg (70). Probes were annotated with NetAffx from the Affymetrix website (*www.affymetrix.com*) and linked to the CAGE and SAGE top 30 genes based on gene symbols.

To annotate the biological processes we took the top 30 differentially regulated genes from CAGE and SAGE (with a Bayesian Error rate $< 1 \times 10^{-50}$ and sorted for differentiated cells on a ratio of differentiated to proliferating cells), as well as the microarray data (sorted on adjusted p-value), and ran these against 7,689 GO (91; 92) Biological Processes in Anni 2.1 (90).

6.3.5 Sequence Annotation

All CAGE and SAGE regions were annotated based on the ElDorado genome annotation (Genomatix, Version 07-2008) for being located in exons, introns, or intergenic regions. Regions that covered an exon and neighboring intron or intergenic region were categorized as partial. In addition a region was categorized as a promoter if it was located in the ElDorado defined promoter region of a transcript. The distance to the nearest TSS (upstream or downstream) was also calculated. CAGE regions were correlated with CAGE data available in ElDorado (originating from the FANTOM3 project (8)).

6.3.6 CAGE Region Confirmation

To confirm that our CAGE regions represented newly discovered 5′-ends of transcripts, we designed primers within CAGE regions upstream of 8 genes (*Bpag*, *Cpeb1*, *Junb*, *Myl1*, *Pik3ca*, *Ppt2*, *Sertad4x*, and *Usp34*, primers in Additional Table 6.1). RT-PCR experiments were performed using random hexamer priming for cDNA synthesis and qPCR performed on a Roche Lightcycler 480. To provide additional validity to these CAGE regions we inspected multiple UCSC tracks (UCSC genes, Ensembl (3) genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome).

To validate that our novel CAGE regions were indicative of myogenic promoters we took all differentially expressed CAGE regions (see Results), expanded or contracted them to a length of 2000 bp, retrieved sequences with Ensembl Perl API scripts, and ran them through CORE_TF (76), a program that identifies over-represented TFBSs. For a background sequence we used 2000 mouse promoters defined as 1000bp before

and 1000bp after the annotated TSS. A Match (55; 51) setting to minimize the sum of false positives and false negatives was used.

We looked into more detail at the upstream CAGE regions of *Myl1*, a myogenic gene that was confirmed to have differential expression in the differentiation analysis. To this we performed standard PCR for a primer set that spans the novel CAGE region into the first UCSC exon (F- TCAGCCAAAATTCCAAGTTGA, R-CCTCCAGAAGAACCTGTCAGA). We also checked this CAGE region, plus 500 bases upstream sequence, for functional evidence. This was done by taking the mouse sequence, searching for orthologous sequences, and identifying conserved patterns of TFBSs, as has been previously described (152; 153).

6.4 Results:

6.4.1 The Biological Model and Experimental Set-up

To study gene expression levels during myogenic differentiation we used C2C12 mouse myoblasts, a common cell model for myogenesis, combined with NGS technology. RNA was isolated from three independent cultures, both of proliferating and differentiated cells. At the latter condition, cells had differentiated into fused and multinucleated myotubes. To confirm successful differentiation, qPCR was performed to determine the expression levels of the genes encoding the late myogenic TF Myogenin and the master myogenic regulator MyoD. Both of these should be expressed at higher levels in differentiated than proliferating cells. qPCR confirmed that cells had started to express Myogenin in differentiated cells and had higher expression of MyoD in differentiated cells (Additional Figure 6.1). CAGE and SAGE libraries were then prepared from all six RNA samples (three independent cell cultures for both proliferating and differentiated cells) and used to determine expression levels based on measurements in the $5'$ and $3'$ region of the transcripts, respectively. We used both methods to evaluate how well transcript level measurements compare and to improve transcript structure annotation. The latter is essential to facilitate bioinformatic approaches to analyze overall transcription regulation based on shared TFBS promoter profiles.

6.4.2 General Sequencing Data and Alignments

Each CAGE and SAGE library was sequenced on a single lane of the Illumina Genome Analyzer II. To investigate technical reproducibility, two CAGE samples (one from proliferating and one from differentiated cells) were sequenced in duplicate. After running the Illumina Genome Analyzer Pipeline for image and sequence quality analysis, we obtained on average 4.5 and 6.9 million reads from the CAGE and SAGE libraries, respectively (Table 6.1). The scarf files, converted to FASTQ format, containing the reads are available at GEO (154) under the accession number GSE21580. We aligned these reads to the repeat masked mouse reference genome and were able to uniquely map (reporting alignments that are unique to one position in the genome), on average, 1.9 million (42%) and 4.1 million (59%) tags for CAGE and SAGE, respectively (Table 6.1).

Table 6.1: Sequencing Results

CAGE sample	# reads sequenced	# reads aligned	percent aligned
Prolif-1	4886341	2086233	42.7%
Prolif-1 duplo	3933233	1770247	45.0%
Prolif-2	5003964	2421443	48.4%
Prolif-3	4734605	2062081	43.6%
Diff-1	4525321	1679081	37.1%
Diff-1 duplo	3101153	1252451	40.4%
Diff-2	5060041	2195263	43.4%
Diff-3	4830194	1578087	32.7%
SAGE sample	# reads sequenced	# reads aligned	percent aligned
Prolif-1	5941753	3351426	56.4%
Prolif-2	7768787	4464057	57.5%
Prolif-3	6723476	3878953	57.7%
Diff-1	9467926	5811947	61.4%
Diff-2	7269002	4618715	63.5%
Diff-3	4392416	2494618	56.8%

Indicators for CAGE and SAGE samples: Prolif for proliferating cells and Diff for differentiating cells, followed by a number representing the biological triplicates. For CAGE there are sequencing duplicates indicated by "duplo." The table contains the number of reads, the number of reads that align uniquely to the repeat masked genome, and the percent aligned.

For visual analysis we constructed UCSC genome browser wiggle files. The wiggle files are available at GEO under accession number GSE21580 and at *http://www.lgtc.nl/publications/Hestand_2010_CAGE_SAGE_wig/*. To retain information on the direction of transcription, there is one file for each strand. In Figure 6.1 we show an example wiggle track for the *Myod1* gene. We clearly see the sharp SAGE peak starting at the most 3′-CATG site followed by 18 additional nucleotides. The CAGE peak at the 5′-end of the transcript is wider, reflecting the variability in the transcription start position. As observed before (8), and observed for many other genes in the current study, CAGE also detects transcription starts in the 3′-region of the gene. This phenomenon is further discussed in the Annotation and Discussion sections. As expected, CAGE and SAGE consistently detect higher expression of *Myod1* in differentiated compared to proliferating cells.

We identified 742,355 CAGE regions, consisting of adjacent nucleotides with aligned reads (after concatenating reads permitting gaps of maximally 100 nucleotides to resolve gaps in alignments due to non-unique genomic sequences). 361,655 SAGE regions (not concatenated, since SAGE tags always start at a fixed position) were identified. After applying a threshold of two tags-per-million, a threshold for very low abundant expression (∼0.3 copies per cell (36)), 41,862 CAGE and 43,512 SAGE regions remained. The CAGE regions have median lengths of 314 nucelotides, and usually represent clusters of TSSs (plus ∼26 nucleotides of downstream sequence). The SAGE tags were 21 or 22 nucleotides long (including the 4 CATG nucleotides representing the NlaIII restriction site).

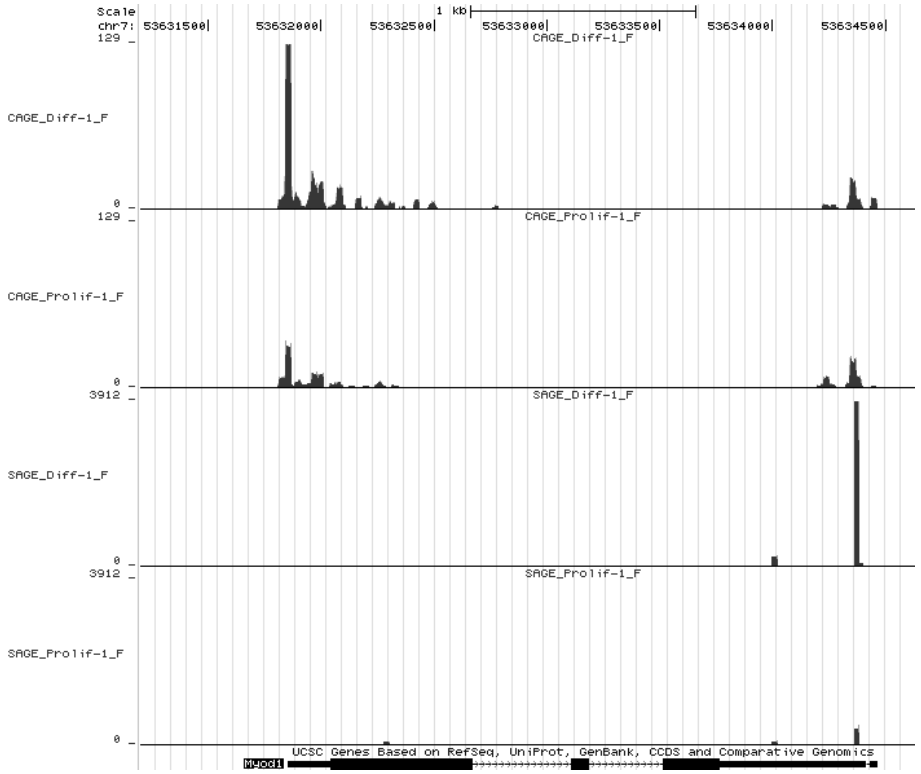


Figure 6.1: CAGE and SAGE wiggle tracks for proliferating (Prolif) and differentiated (Diff) cells in the UCSC Genome Browser for the myogenic marker *Myod1*. We only display reads aligning to the forward strand, the coding direction for *Myod1*. Chromosomal positions are indicated at the top. For each track the Y-axis scale corresponds to the number of tags aligned at that genomic position. Scales use a maximum from each relevant technique in this viewing window (129 for CAGE and 3912 for SAGE). There is 5′ and 3′ concordance for CAGE and SAGE samples, respectively. CAGE provides broader peaks, reflecting TSSs plus ∼26 nucleotides of downstream sequence, wheres SAGE provides discrete peaks. A higher number of tags are in differentiated compared to proliferating samples.

6.4.3 Technical Reproducibility and Biological Overlap

A high correlation was found between the technical CAGE replicates (median Pearson correlation of 0.981) as well as the biological triplicates (median Pearson correlation of 0.963 (Figure 6.2A/B, Additional Table 6.2). As expected, correlation between proliferating and differentiated cells was lower (median Pearson correlation of 0.771) (Figure 6.2C, Additional Table 6.2). Similarly, we observed a high reproducibility for the SAGE experiments (median Pearson correlation of 0.930) between biological triplicates (Figure 6.2D and Additional Table 6.2). Again, the correlation between proliferating and differentiated cells (median Pearson correlation of 0.839) was lower than between cells from the same condition (Figure 6.2E and Additional Table 6.2).

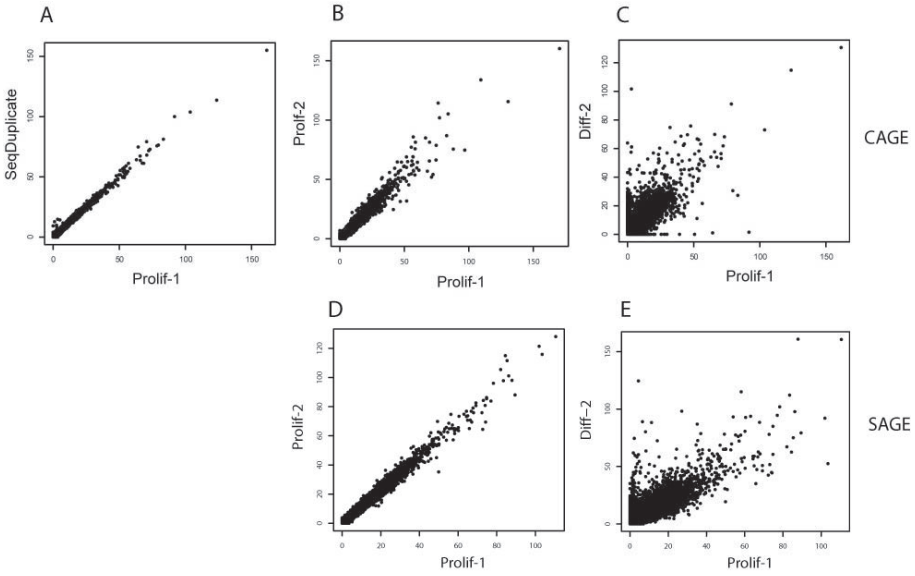


Figure 6.2: High reproducibility was found in CAGE regions between sequencing duplicates (A) and biological replicates (B). Panel C shows correlation between CAGE samples from proliferating and differentiated cells. High reproducibility can also be found between SAGE biological replicates (D). Panel E shows the correlation between CAGE samples from proliferating and differentiated cells. The plotted values represent the square root of the number of tags per region.

6.4.4 Annotation of Regions

We annotated the 41,862 CAGE regions using Eldorado's mouse genome annotation: 9,957 regions map to an annotated exon, 27,190 partially overlap an exon and intron/intergenic region, 2,368 map to an intron, and 2,347 regions are purely intergenic. The median number of tags in the exonic and partial regions (63 tags and 90 tags respectively) were higher than in the intronic and intergenic regions (45 tags and 54 tags, respectively). These data clearly show that our CAGE experiments identifies many (lower abundant) TSSs / transcribed regions that have not yet been identified and/or annotated as such in current genome databases.

Based on ElDorado annotation of our 41,862 CAGE regions, 13,541 of the CAGE regions (32%) contained an annotated TSS, 6,331 CAGE regions (15%) were annotated as promoters (i.e. a genomic region surrounding a TSS containing functional elements like TFBSs that are responsible for the regulation of the expression of the transcript), and 8,028 (19%) CAGE regions contained an annotated transcript 3′-end. 3′-end alignments are consistent with the previously observed (8) significant amount of (shorter) transcripts originating from the 3′-ends of genes. We compared our CAGE results to previous CAGE studies (FANTOM3) contained in ElDorado and identified 31,680 regions (76%) overlapping with at least one on the FANTOM3 CAGE tags. Only 6,119 (15%) and 5,635 (13%) of these regions were observed in FANTOM3 muscle and heart CAGE libraries, respectively. This is explained by the

small size of these muscle and heart libraries (8), together representing only 1% of all available CAGE tags in FANTOM3.

6.4.5 Comparison of CAGE, SAGE, and Microarray Expression Data

To compare overall expression level measurements we assigned CAGE and SAGE regions to genes (including 1000 bases upstream and downstream of the gene). Expression above a threshold of 2 transcripts per million (∼0.3 copies per cell) (155) was observed for 10,409 and 10,987 genes respectively. Expression profiles for both methods showed a high correlation (Figure 6.3A-C), with 9,240 genes being expressed in both methods above 2 transcripts per million (Figure 6.3D). Additional Figure 6.2 shows that the relative overlap is even bigger when higher detection thresholds are applied, obviously at the expense of many more genes not reaching the detection threshold. 4,304 genes were differentially expressed between proliferating and differentiated cells (Bayesian error rate < 0.05) according to the CAGE data and 3,846 according to the SAGE data with 2,144 genes present in both lists of significant genes (Figure 6.3E). Most others were just borderline significant according to one of both methods.

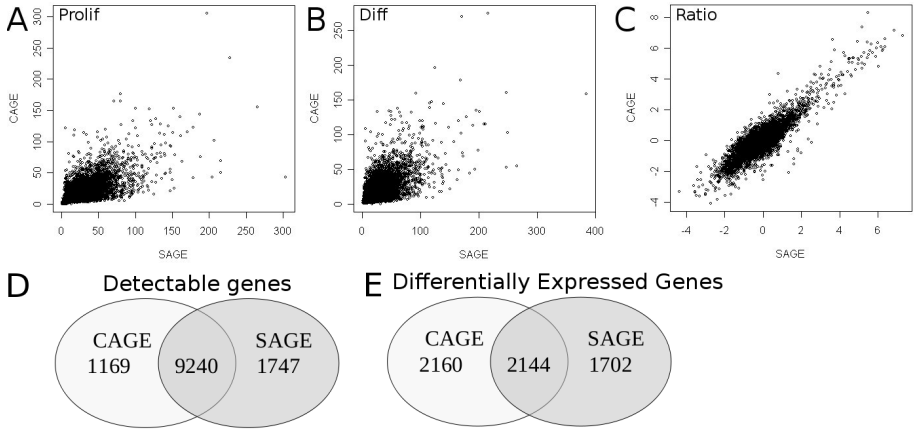


Figure 6.3: Correlation of CAGE versus SAGE for proliferating samples (A), differentiated samples (B), and the ratio of proliferating / differentiated cells (C). Values are the square root of the number of tags per gene for A and B. For C the values are the log ratio of the normalized number of tags per gene in differentiated over proliferating cells. The overlap of detectable genes (D) and differentially expressed genes (E) between CAGE and SAGE is indicated.

We compared the top 30 most differentially expressed genes for both methods (Table 6.2A) to results from a similar microarray dataset on myogenic differentiation in the same cell line (148). In general, the genes identified by CAGE and SAGE also demonstrated very significant changes on the microarrays. However, in the top 30, 13 genes identified by CAGE and 10 identified by SAGE were not represented on the array, demonstrating the comprehensive nature of the CAGE and SAGE-

based gene expression profiling techniques. The biological processes controlled by the top 30 CAGE, SAGE, and microarray genes, were annotated with the Anni2.1 text-mining tool (Table 6.2B). All CAGE and SAGE-derived GO terms can readily be related to muscle development, whereas 3/10 GO terms associated with the microarray-derived gene list can not ("cyclin-dependent protein kinase inhibitor activity," "6-phosphofructokinase," and "tumor suppressor activity").

6.4.6 Differential TSS Use and Validation

In our CAGE data, we identified 111 regions upstream of the start of a known gene and 85 CAGE regions downstream of an annotated gene containing significantly different numbers of tags in proliferating and differentiated cells (Additional Table 6.3). The differential expression of transcripts originating from 7 out of 8 of these regions (upstream from genes *Bpag*, *Cpeb1*, *Junb*, *Myl1*, *Pik3ca*, *Ppt2*, *Sertad4x*, and *Usp34*) were confirmed by RT-PCR/qPCR (Figure 6.4B and Additional Figure 6.3). To evaluate if these novel exons were contained in a transcript of the gene of interest we inspected the following tracks in he UCSC genome browser: UCSC genes, Ensembl genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome (Figure 6.4A and Additional Figure 6.4). In all but *Junb* we found the CAGE regions overlapping at least one exon from an additional track connected to the gene of interest (Figure 6.4A and Additional Figure 6.4). This indicates that these CAGE regions usually represented alternative transcripts that are not yet properly annotated in all resources, including the mainstream UCSC and Ensembl annotations. This suggests that the mainstream genome annotation are far from complete and that additional evidence, including our CAGE data, is required to more precisely define transcript structure.

To support that differential transcription in the 196 CAGE regions is regulated by myogenic TFs, we searched for over-represented TFBSs and found the binding sites for the master regulators MyoD (p-value 6.49×10^{-03} from CORE_TF's binomial test) and Myogenin (p-value: 3.87×10^{-02}) and the Ebox motif (p-value 6.02×10^{-03}) (frequently found in muscle promoters (156; 157)) to be significantly over-represented in 2,000 bp of sequence composed of the CAGE and surrounding regions (Additional Table 6.4).

For one of these novel CAGE regions, *Myl1*, we confirmed by standard RT-PCR that there is a transcript extending from the novel CAGE region into the UCSC defined exon 1 (Figure 6.4C). The CAGE sequencing, RT-PCR/qPCR within the region, and the standard PCR into exon 1 all confirmed that this transcript is only present in differentiated cells, explaining why it is missing in standard genome annotations. For functional evidence that this region is used as a promoter, we also looked for conserved TFBSs in and upstream of this region. Within the Genomatix Suite we identified orthologous sequence regions from human and horse corresponding to the CAGE region and $5'$ upstream (promoter) sequence. In this area we identified conserved TFBSs for NKX, GATA, and SRF (Figure 6.4D), all of which are known to be involved in the regulation of muscle genes (158). This makes it likely that the region directly upstream of the novel exon 1 is used as an alternative promoter.

Table 6.2: Differential Gene Expression

A.CAGE Gene	Ratio	Microarray p-val	SAGE Gene	Ratio	Microarray p-val
Hfe2	4073	NA	RP23-36P22.5	576	NA
Myom3	1624	NA	Neb	525	NA
Lmod2	1305	NA	Mylpf	504	1.70×10^{-15}
Myh7	1124	5.98×10^{-3}	Ttn	380	NA
Mb	908	1.07×10^{-14}	Myh3	368	2.40×10^{-14}
RP23-36P22.5	735	NA	Xirp1	306	2.24×10^{-13}
Pygm	717	4.82×10^{-17}	1110002H13Rik	263	NA
Myl4	614	8.86×10^{-20}	Tnnc1	232	1.24×10^{-11}
Synpo2l	595	NA	Cav3	150	3.58×10^{-22}
Myh1	561	3.64×10^{-15}	Cbfa2t3	133	2.89×10^{-10}
Tnni1	529	2.24×10^{-9}	Chrng	115	4.63×10^{-9}
Tnni2	442	3.20×10^{-11}	Myom2	105	6.66×10^{-16}
Mpa2l	410	NA	Tnnt1	100	1.15×10^{-10}
Ctrb1	406	7.55×10^{-7}	Ryr1	92	7.03×10^{-14}
Ttn	402	NA	Apobec2	84	2.95×10^{-15}
Neb	374	NA	Cox6a2	72	2.45×10^{-16}
Kcnq4	365	NA	Dio2	64	2.14×10^{-10}
Mylpf	341	1.70×10^{-15}	C1qtnf3	52	4.36×10^{-5}
1110002H13Rik	341	NA	Htr2b	43	3.76×10^{-6}
Inpp4b	328	NA	Sgcg	42	1.15×10^{-12}
Xirp1	307	2.24×10^{-13}	Fndc5	39	NA
Atp2a1	304	2.06×10^{-14}	Jsrp1	36	NA
Casq2	297	4.74×10^{-6}	Ankrd23	36	NA
Cacna1s	296	5.20×10^{-19}	AK031267	29	NA
Ces2	245	NA	Sema6a	26	3.08×10^{-3}
Cox6a2	241	2.45×10^{-16}	Lgr5	23	9.33×10^{-1}
Myog	238	2.36×10^{-6}	Pdlim3	22	3.18×10^{-6}
Myh3	234	2.40×10^{-14}	Klhl31	22	NA
Tmem182	216	NA	ORF63	21	NA
Tnnc1	215	1.24×10^{-11}	Gfra2	19	2.98×10^{-2}

B.CAGE GO	SAGE GO	microarray GO
1)regulation of striated muscle contraction	1)regulation of muscle contraction	1)*cyclin-dependent protein kinase inhibitor activity*
2)cardiac muscle contraction	2)cardiac muscle contraction	2)Myogenesis
3)Myogenesis	3)Myogenesis	3)skeletal muscle development
4)regulation of muscle contraction	4)regulation of striated muscle contraction	4)myoblast differentiation
5)skeletal muscle development	5)skeletal muscle development	5)*6-phosphofructokinase activity*
6)Muscle Development	6)myofibril assembly	6)Muscle Development
7)striated muscle contraction	7)Muscle Development	7)muscle cell differentiation
8)myoblast differentiation	8)myoblast fusion	8)*tumor suppressor activity*
9)muscle cell differentiation	9)striated muscle contraction	9)myofibril assembly
10)sarcomere organization	10)muscle cell differentiation	10)heart development

Top 30 genes from SAGE and CAGE expression data (A). All genes with a Bayesian error rate $< 1 \times 10^{-50}$ were sorted on the ratio (normalized tags from differentiated / proliferating cells) and the highest ratios for differentiated cells displayed. The microarray p-values are adjusted p-values for differential gene expression from a similar experiment (proliferating and differentiated C2C12 cells (148)). NA = no probe annotation for the gene. The top 10 GO biological processes (B) associated with the top 30 genes for CAGE, SAGE, and microarray experiments indicate clear muscle relations, with the exception of 3 (in italics) processes in the microarray data.

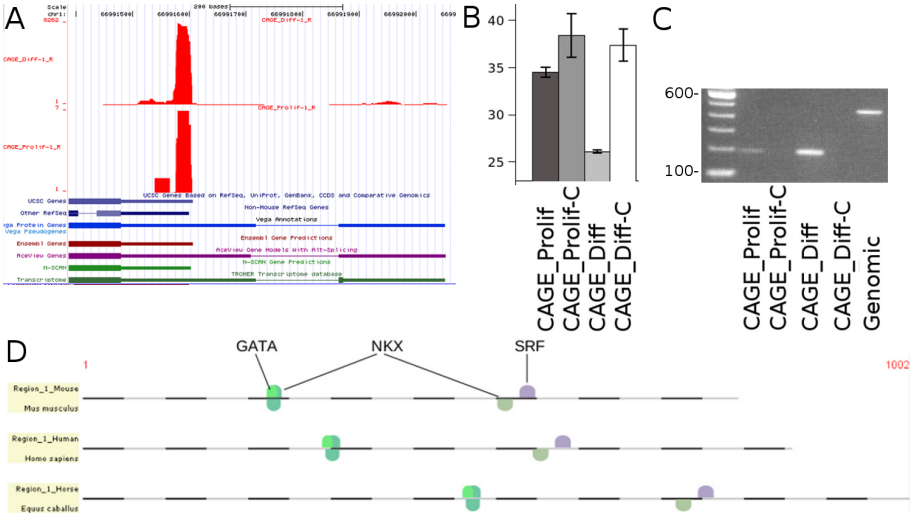


Figure 6.4: The UCSC display of (A) UCSC/Ensembl defined first exon and an upstream CAGE region for *Myl1* (reverse strand reads only, on which the gene lies) for samples Prolif-1 and Diff-1. The Y-axis indicates the number of tags aligned at each position in the genome. We also display additional track information (UCSC genes, Ensembl genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome), several of which confirm the presence of the upstream CAGE region. (B) qPCR with primers within the CAGE region for Prolif, Prolif-C (reverse transcriptase control), Diff, and Diff-C (reverse transcriptase control). The qPCR results are plotted as threshold cycle (Cp) values (lower = higher expression), with bars indicating a range of one standard deviation between technical duplicates. (C) standard PCR on agarose gel with forward primer in the novel CAGE region and reverse primer in the conventional exon 1. Comparison with the genomic control verifies the presence of an intron of 200 bases. A 100 bp ladder is included. Panels A-C are all consistent with higher expression in differentiated than proliferating cells. (D) Cross-species conserved muscle specific TFBSs around and upstream of the *Myl1* CAGE region support its role as a promoter for this region.

6.5 Discussion

Using CAGE and SAGE methods with NGS we have measured gene expression levels during myogenic differentiation and identified muscle specific TSSs. By elucidating promoter regions and regulation in these myogenic cells we hope to better understand the process of muscle development and regeneration, providing clues to cure muscle related illnesses. Since biologists and clinicians often study (first) exons and 5′ promoter regions it is crucial to know the positions of TSSs in the genome. Our data will help them identify potentially pathogenic mutations in transcripts and promoters used during myogenic differentiation, which might have been over looked with current genome annotations. On a technical level, this is the first time CAGE and SAGE have been evaluated using the same RNA samples.

We found both the technically demanding CAGE method and the slightly less

laborious SAGE method to be extremely robust. Biological triplicates with independent sample preparations and sequencing runs were found to have high correlations (Figure 6.2, Additional Table 6.2). This is in line with previous findings in the FANTOM4 CAGE study (159) and our previous (36) finding with SAGE. Higher technical reproducibility also enhances the ability to verify low expressed genes, which was an obstacle in microarray analysis. The high quality of the data implies that more investments should be made in biological than technical replicates, as demonstrated for CAGE for the first time in the current paper.

This study also highlights other advantages over microarrays. For a third of the top 30 genes (13/31 CAGE genes and 10/30 SAGE genes, Table 6.2A) there was no probe on the microarray. Finding many more significant genes not interrogated by the microarrays stresses the more comprehensive transcript profiling by NGS based methods. We also found more muscle related biological processes associated with the top 30 CAGE and SAGE genes compared to the microarray top 30 genes (Table 6.2B) indicating the higher relevance of the top hits for the process under study.

The data provided by these methods have greatly expanded our knowledge of muscle specific transcription. 56% of the analyzed CAGE regions contained an annotated TSS, indicating discovery of many novel TSSs. 76% of CAGE regions matched known FANTOM3 CAGE tags, but less than 20% of those matched known muscle related CAGE tags. This is likely due to the lower sequencing depth in the previous FANTOM3 CAGE studies. High overlap with the previous FANTOM CAGE regions indicates these reflect true TSSs, but there is a lack of information on the definition of TSS usage in relation to tissue. To exemplify this point, we identified 196 intergenic regions significantly different between proliferating and differentiated cells, indicating muscle-specific alternative promoter and first exon usage. Several of these were verified by PCR and additional UCSC track evidence. We also identified over-represented muscle specific TFBSs in the 196 CAGE regions and additional conserved muscle specific TFBSs upstream of a novel first exon of the *Myl1* gene, coding for one of the light chains of the myosin protein complex involved in muscle contraction. These muscle specific TFBSs indicate that the identified regions potentially serve as a promoters.

This is the first study to compare NGS of CAGE and SAGE libraries from the same RNA samples. Gene expression measurements by CAGE and SAGE are generally consistent. The high correlation between methods (Figures 6.3A-C), large overlap between genes detected (Figure 6.3D) and differential gene lists (Figure 6.3E), and gene involvement in similar biological pathways (Table 6.2B) indicates these methods are interchangeable for expression analysis. Only when transcript structure ($5'$ or $3'$) is important is one method preferential over another. Correct $5'$ usage is crucial for promoter based regulation studies, whereas proper $3'$ usage is needed for studies concerning micro-RNA regulation.

Of the 4,304 and 3,846 genes differentially expressed between proliferating and differentiated cells with CAGE and SAGE, respectively, over half (2,144) of the genes are identical. More changes in CAGE than SAGE levels could indicate alternative promoter usage is more common than that of alternative $3'$ ends. The detection of genes by one technique, but not the other, is mostly inherent to the use of thresholds. In addition, a minority of transcripts may be missed entirely by one of the methods due to the absence of a CATG site in the transcript (SAGE) or the sequence around

the TSS not being unique in the genome (CAGE). Rmap does not report a read when it aligns with equal mismatches to multiple regions in the genome. Therefore non-unique TSS sequences will not be reported and included in our analysis. For both techniques, we frequently detected multiple regions in the same gene. 75% of the genes had multiple SAGE tags with abundance above the threshold of 2 transcripts per million. In our previous paper (36), we discussed that this is probably not a technical artifact but most likely due to different 3′ ends and usage of multiple polyadenylation sites.

Similar to previous studies(8), we found a large number of CAGE tags aligning to the 3′ end of known transcripts. This phenomenon has been previously validated by the RACE method and explained as potential 3′ derived regulatory noncoding RNAs (8). With additional analysis this could serve a a method for identification of noncoding RNAs. In addition, these should be recognized as a source of false expression levels identified by the 3′ based SAGE method and microarrays based on 3′ probes.

67% of the genes contained multiple CAGE regions. This phenomenon was previously referred to as "exon painting" (160). Examples of genes where nearly all exons are covered by CAGE tags are *Col1a1* and *Col1a2* (Additional Figure 6.5A/B, respectively). This is unexpected since the RNA integrity was high in all samples, the CAGE technique only captures capped transcripts, and even when some non-capped transcripts may be included, the method will only create tags from the ultimate 5′ end. Together with the observation of genes with a highly abundant peak at the 5′ end without any exon painting (Additional Figure 6.5C/D) and the fact that the exon painting patterns are highly reproducible in independent CAGE sample preparations, this suggests that there is a biological explanation for the exon painting phenomenon. The observation of exon painting is consistent with the finding of many short transcripts from exonic regions in a tiling array study (160). It is not clear whether these short transcripts are degradation products from larger transcripts, true *de novo* transcriptional events, or a combination of both. From our study, it is highly likely that many of these shorter transcripts contain a cap structure. The process of recapping of transcript fragments has been documented before (160). Fejes-Toth *et al.* propose long RNAs are spliced into mature and translatable RNAs, but that these mature RNAs can also be further processed (160). This further processing involves cleavage into smaller RNA fragments and possible modification by additional 5′ capping (160). The presence of exon painting complicates the identification of novel TSSs and is the reason why we focused on the discovery of novel TSSs in intergenic regions and did not report alternative TSSs within annotated genes. A positive consequence of the exon painting phenomenon is that the CAGE technique gives additional information on the exon structure of many genes.

The large data yield and reproducibility should serve as an example of the advantages of applying NGS to CAGE and SAGE techniques. These methodologies should be expanded to other tissues and processes in the future to enrich our knowledge of the genome of many organisms. This work has provided a substantial increase in our knowledge of myogenic TSSs and expression. This has also demonstrated the technical advantages of CAGE and SAGE in conjunction with NGS.

6.6 Acknowledgments

We wish to thank Michel Villerius, Michiel van Galen, and Ivo Fokkema for their computational assistance. We also wish to thank Rolf Vossen for advice with the Lightcycler 480. We would also like to thank Henk P.J. Buermans and Emile J. de Meijer for wet-lab assistance and providing primers for expression analysis.

6.7 Funding

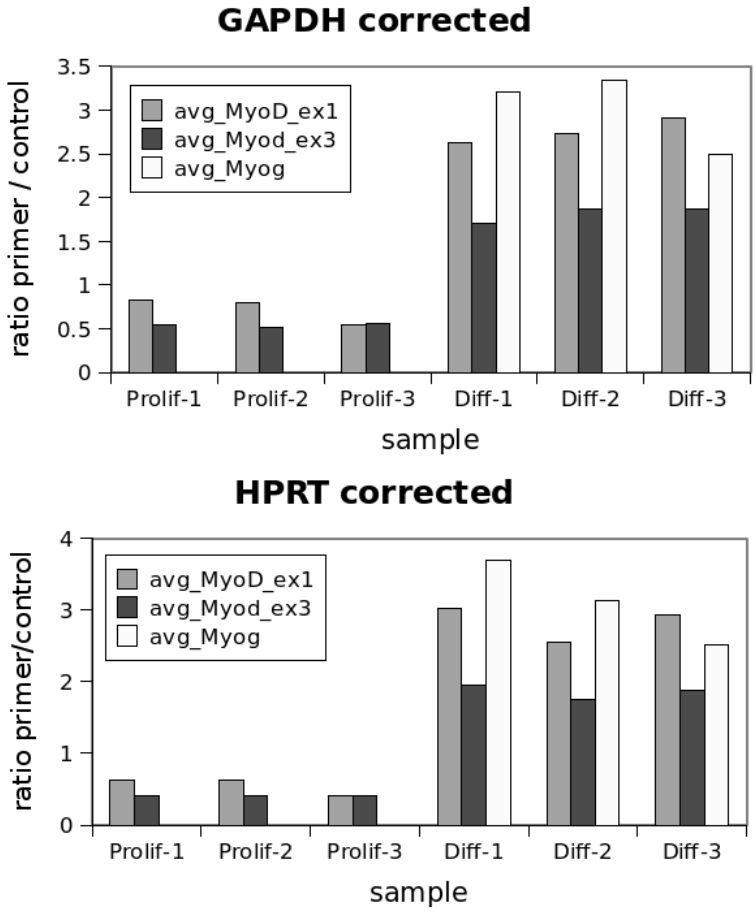
This project was supported by grants from the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) and the Center for Biomedical Genetics (in the Netherlands).

6.8 Conflict of interest

Andreas Klingenhoff, Matthias Scherf, Wilbert van Workum, Makoto Suzuki, Thomas Werner, and Matthias Harbers declare that they have competing financial interests.

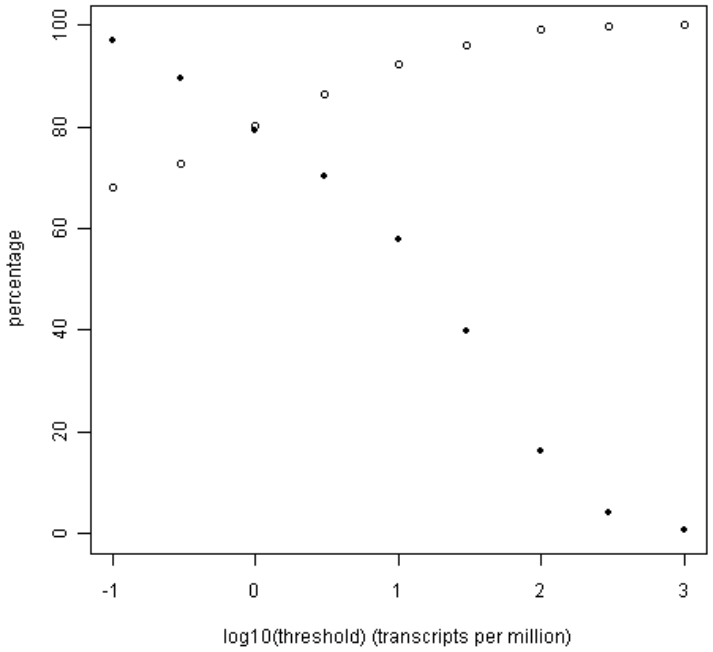
6.9 Additional Files

Additional Figure 6.1



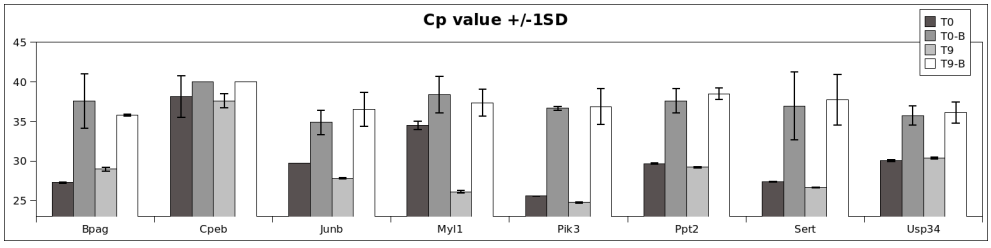
Myogenic confirmation of RNA. Expression levels are relative to the control genes *GAPDH* or *HPRT*.

Additional Figure 6.2



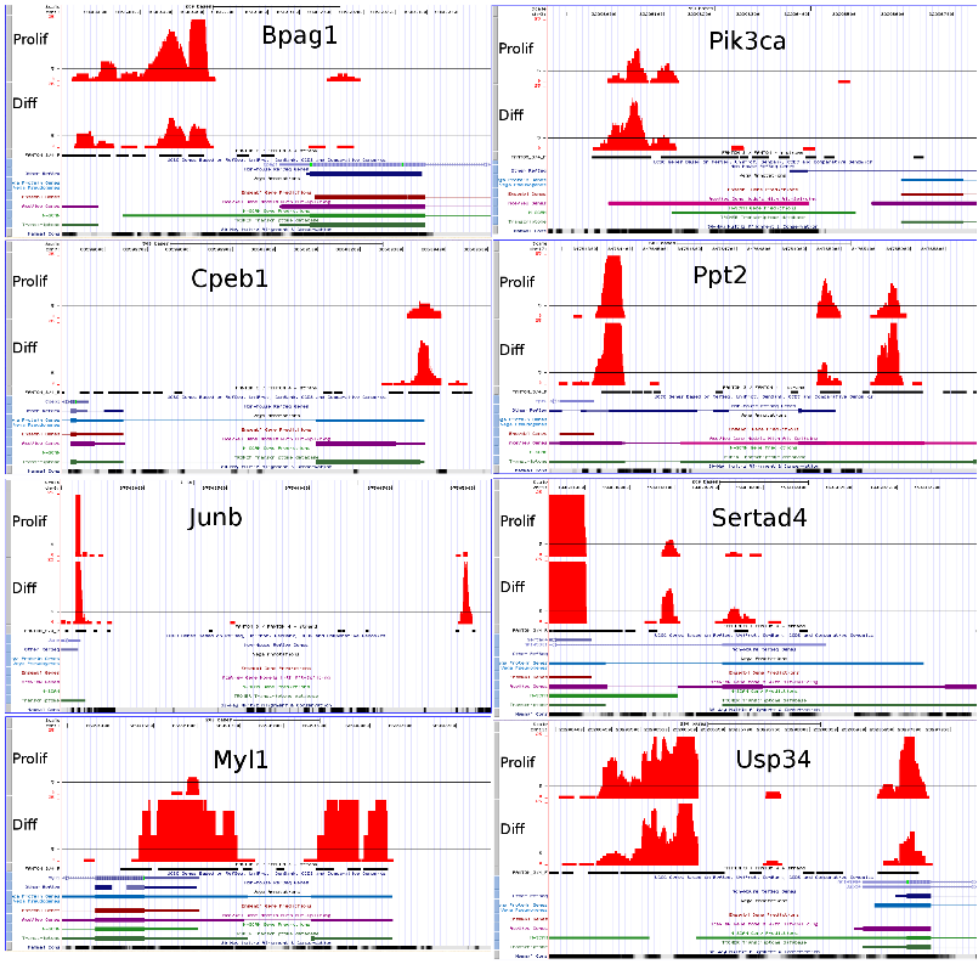
Overlap between genes detected by SAGE and CAGE. We compared the genes detected by SAGE at different threshold values (expression levels in transcripts per million, x-axis, 10log scale) with the genes detected in CAGE with a fixed threshold of 2 transcripts per million. The number of genes also detected by CAGE is expressed as a percentage of the total number of genes detected by SAGE at different thresholds is plotted in open symbols. The closed symbols represent the percentage of genes remaining after thresholding of the SAGE data.

Additional Figure 6.3



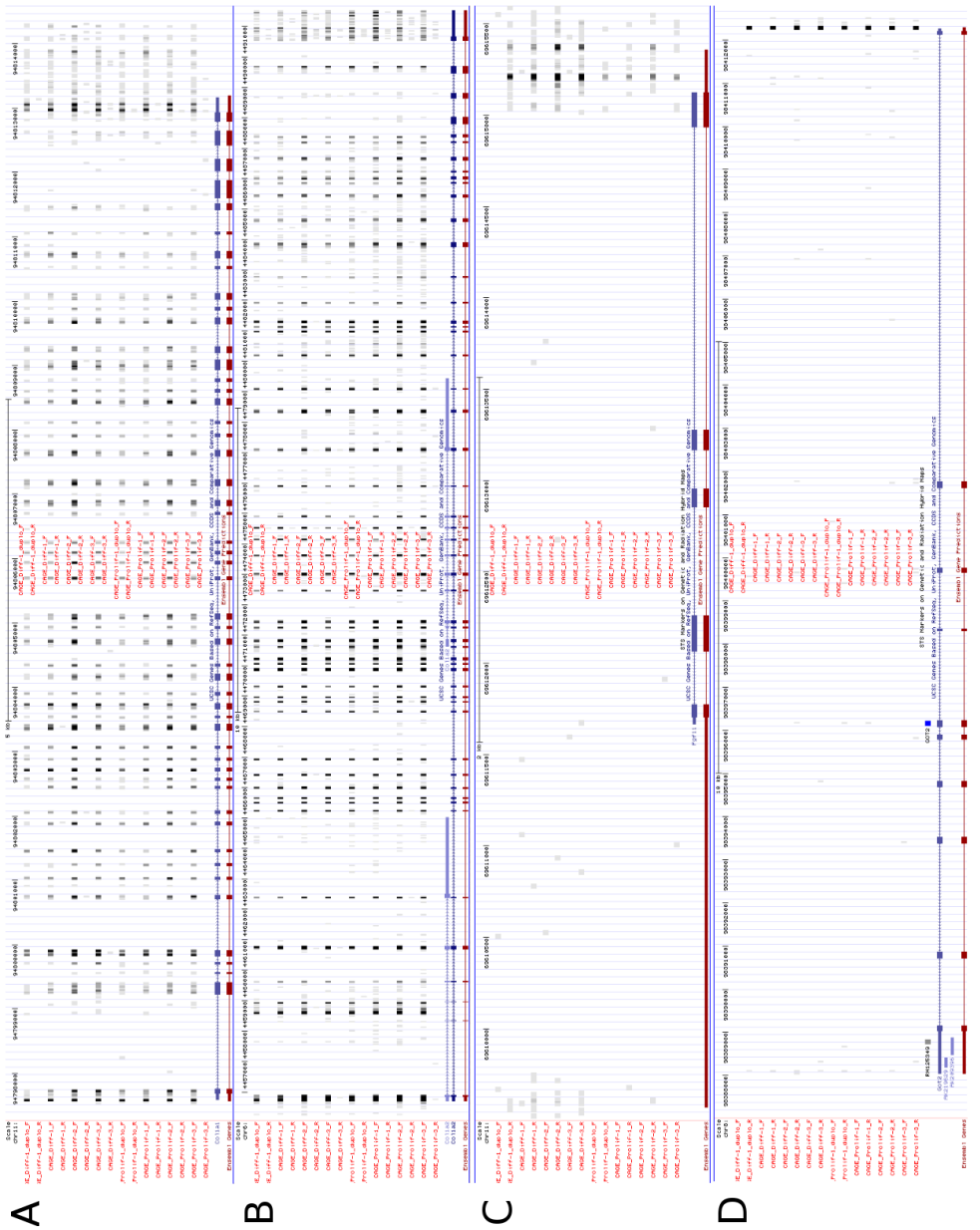
PCR validation of CAGE tags in novel upstream exons. -B in the legend indicates a no enzyme reverse transcriptase control. The Y-axis is the threshold cycle (Cp) value (lower = higher expression), with plotted bars indicating +/- one standard deviation.

Additional Figure 6.4



UCSC/Ensembl defined first exon and an upstream CAGE region for 8 example genes (only reads on the strand which the genes lie on are displayed) for samples proliferating-1 and differentiated-1. The Y-axis indicates the number of tags aligned at each position in the genome. We also display additional track information (UCSC genes, Ensembl genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome), several of which confirm the presence of the upstream CAGE regions. A larger color figure is available upon request

Additional Figure 6.5



Examples of exon painting conservation in 8 different CAGE datasets divided over each strand (F or R) across the genes *Col1a1* (A) and *Col1a2* (B). The same datasets are also plotted against the genes *Fgf11* (C) and *Got2* (D), which do not show exon painting. To ensure regions with a low number of tags are visible and not over-shadowed by regions with a high number of tags the maximum display is set to 25 tags per position.

Additional Table 6.1

Expression	
Primer Name	Sequence
MyoD_ex3-F	CCCAATGCGATTTATCAGGT
MyoD_ex3-R	TCTGCTCTTCCCTTCCCTCT
MyoD_ex1-F	GACAGGGAGGAGGGGTAGAG
MyoD_ex1-R	AAGTCTATGTCCCGGAGTGG
MyoG-R	TGGGAGTTGCATTCACTGG
MyoG-F	CCTTGCTCAGCTCCCTCA
HPRT-F	TCCCTGGTTAAGCAGTACAGCC
HPRT-R	CGAGAGGTCCTTTTCACCAGC
GAPDH-F	TCCATGACAACTTTGGCATTG
GAPDH-R	TCACGCCACAGCTTTCCA
CAGE region	
Primer Name	Sequence
Myl1_CF	TCAGCCAAAATTCCAAGTTGA
Myl1_CR	CCACTTCCTAAGAAGCTTTACCG
Usp34_CF	CGGACGGAAGAGGAAAGG
Usp34_CR	GCCTCTCTCCGCACACAC
Ppt2_CF	CACTGGCAGGGTTTGTGTC
Ppt2_CR	GACAAACTGCTCCTCAGATCC
Bpag_CF	GTGCTGAGTCATGGCGAGAG
Bpag_CR	CCGGAACGACTGATGGAG
Pik3_CF	GTGGGGAAGAGTTCGTTGTTT
Pik3_CR	GGTCTCTCTTTCCGCTCACAT
Cpeb_CF	GTCTGGTCCAGCCCTAGC
Cpeb_CR	GAAGCTGTTGTTCCGAGAGG
Sert_CF	GCTCAGTCCAGCTCTGACATC
Sert_CR	CTTCCCCTCTGTACAGCACAC
Junb_CF	GGAAGGAGGACTTAAGGGTCA
Junb_CR	GTAGGGGCATTGGAGAAGAAG

PCR primers used in study.

Additional Table 6.2

CAGE	Prolif-1^d	Prolif-1	Prolif-2	Prolif-3	Diff-1^d	Diff-1	Diff-2	Diff-3
Prolif-1^d	1.000	0.983	0.960	0.959	0.754	0.754	0.787	0.770
Prolif-1	0.983	1.000	0.961	0.959	0.751	0.755	0.790	0.771
Prolif-2	0.960	0.961	1.000	0.982	0.771	0.774	0.810	0.783
Prolif-3	0.959	0.959	0.982	1.000	0.765	0.767	0.804	0.777
Diff-1^d	0.754	0.751	0.771	0.765	1.000	0.978	0.963	0.962
Diff-1	0.754	0.755	0.774	0.767	0.978	1.000	0.970	0.967
Diff-2	0.787	0.790	0.810	0.804	0.963	0.970	1.000	0.966
Diff-3	0.770	0.771	0.783	0.777	0.962	0.967	0.966	1.000

SAGE	Prolif-1	Prolif-2	Prolif-3	Diff-1	Diff-2	Diff-3
Prolif-1	1.000	0.968	0.920	0.806	0.879	0.839
Prolif-2	0.968	1.000	0.940	0.802	0.885	0.851
Prolif-3	0.920	0.940	1.000	0.721	0.852	0.838
Diff-1	0.806	0.802	0.721	1.000	0.887	0.824
Diff-2	0.879	0.885	0.852	0.887	1.000	0.941
Diff-3	0.839	0.851	0.838	0.824	0.941	1.000

Reproducibility of CAGE and SAGE methods: Pearson correlations. ^{d} = sequencing duplicate.

Additional Table 6.3

available upon request

Regions upstream and downstream of a known genes that were significantly different between proliferating and differentiated cells. Distance is distance from the CAGE region to the TSS. Bay.error is the Bayesian error rate.

Additional Table 6.4

TFBS	p-value
V$SREBP1_Q6	6.9010×10^{-04}
V$BACH2_01	3.6910×10^{-03}
V$CREB_01	5.7810×10^{-03}
V$EBOX_Q6_01	6.0210×10^{-03}
V$BACH1_01	6.2110×10^{-03}
V$MYOD_01	6.4910×10^{-03}
V$PTF1BETA_Q6	8.0810×10^{-03}
V$PR_01	1.0110×10^{-02}
V$NRSE_B	1.0610×10^{-02}
V$TFE_Q6	1.1210×10^{-02}
V$USF_C	1.1710×10^{-02}
V$E12_Q6	1.2110×10^{-02}
V$SZF11_01	1.4010×10^{-02}
V$AML1_01	1.4510×10^{-02}
V$CACBINDINGPROTEIN_Q6	1.4610×10^{-02}
V$SREBP_Q3	1.6310×10^{-02}
V$TGIF_01	1.6910×10^{-02}
V$FXR_IR1_Q6	1.7110×10^{-02}
V$AP4_01	1.7210×10^{-02}
V$PADS_C	1.8410×10^{-02}
V$STRA13_01	2.1910×10^{-02}
V$AREB6_03	2.9110×10^{-02}
V$RREB1_01	3.0710×10^{-02}
V$NFY_C	3.3510×10^{-02}
V$ETS1_B	3.3910×10^{-02}
V$HTF_01	3.6310×10^{-02}
V$NANOG_01	3.7510×10^{-02}
V$MYOGENIN_Q6	3.8710×10^{-02}
V$ZTA_Q2	4.5110×10^{-02}
V$COUP_DR1_Q6	4.6510×10^{-02}

Top 30 (sorted on decreasing p-value significance) CORE_TF over-represented TFBSs (represented as a TRANSFAC position weight matrix) in the 196 differentially expressed intergenic CAGE regions.

Chapter 7

Discussion

This thesis presents dry and wet lab techniques to elucidate the involvement of transcription factors (TFs) in the regulation of the cell cycle and myogenesis. However, the techniques described in this manuscript could be used for the study of other TFs in these and other biological processes. These two methodologies complement each other. *In silico* analysis provides clues into what to verify in the wet lab, which can be used as the basis for additional *in silico* predictions. When developing genetics and genomics to study complex cellular processes these methodologies are essential to successfully tackle the large quantities and complexities of data successfully.

7.1 *In Silico* Prediction of Transcription Factor Binding Sites: Past, Present, and Future

Figure 7.1: PWM evolution: The sequence affinity of TFBSs has evolved from single sequences, to PWMs, to larger and larger databases of PWMs.

Computational predictions of TF binding sites (TFBSs) have come a long way. From initial single specific sequences, such as myogenic regulatory factors binding an E-box (simply the sequence CANNTG, reviewed in Sabourin *et al.* 2000 (161)), the jump was made to position weight matrices (PWMs), accounting for the variation in sequences bound by specific TFs (Figure 7.1). PWMs have also accounted for combinations of TFs serving as complexes, such as the TRANSFAC PWM V$MYOGNF1_01 for Myogenin and NF1 (162). As more PWMs become available (such as in a database) we can mine sequences for multiple PWMs indicating co-regulation and competition between multiple TFs.

However, since at present there are only a few hundred PWMs and the human proteome is estimated to contain approximately 2600 proteins with DNA binding

domains (163), there is still a lot to discover. Existing methods for obtaining experimental based PWMs include analysis of ChIP-chip and ChIP-seq data. With the cost of ChIP-seq decreasing and the popularity increasing there are more and more sequences defined as TF targets, for which we can extrapolate motifs using programs like MEME (45; 46) and Gibbs samplers (47; 48; 49). The major limitation in obtaining PWMs based on ChIP-seq data is becoming the availability of good antibodies, placing the bottleneck more on the biology and less on the informatics. Hopefully in the near future we will have a greater number of PWMs, based on a larger quantity of data resulting in higher quality.

The use of a PWM to predict TFBSs has also evolved over time. In the past programs like Match (51; 55) could identify whether a single PWM matched a sequence or not, given a similarity over a given threshold. This could be done in a batch setting for multiple PWMs, but calculations were still done one TF at a time. This program is still useful today due to its speed, but has several theoretical limitations. Three limitations are using a threshold, not accounting for competition, and not considering the TF concentration in a cell. All three of these principles are accounted for in Sunflower (56). Sunflower does not report back a black or white, bound or not bound, report, but instead a probability of a TF binding a sequence. Therefore, this also models the binding affinity of a TF. The Sunflower algorithm, by nature, introduces competition between PWMs for the same nucleotide sequence. We showed improved results when looking for enrichment of TFBSs in sequences with Sunflower compared to Match in chapter 3. In addition, though the current model sets the concentration of all TFs equal by default, they can be adjusted on a TF by TF basis. In the future, models will take into account additional factors that contribute to the binding and functionality of TFs, including TF concentration, chromatin state, and methylated nucleotides. The first steps have been made in this direction: *e.g.* Segal et al. (81) have included TF concentrations in their model to identify TFBSs in Drosophila.

One issue, addressed in chapters 2 and 3, is the use of proper background sequences when looking for enrichment in a foreground set of sequences. One argument is that background sets should have similar properties as the foreground set to identify TFBSs properly. For example, when a foreground set has a high TA content compared to the background set there is a high likelihood that TATA binding proteins will be predicted. Therefore, TATA box proteins predicted may well be false positives. However, TFs have access to all sequences in the genome so some may argue that it is improper to make these selections. In our searches for MyoD and Myog in expression data we found that matching foreground and background promoters based on GC content improved results. Incorporating GC content into predictions for *de novo* motifs has also been shown to improve results (86; 87). Since CpG islands have a higher GC content by definition (10) and potentially different promoter binding behavior (8; 9) it is also appropriate to sort data on CpG content. Besides sorting on GC and CpG content, we investigated, in chapter 3 with ChIP-chip and ChIP-PET data, sorting on presumed better annotated promoters (containing a $5'$ UTR in Ensembl). However, we found that this did yield better results. In the same chapter we also compared the use of random genomic sequence as background instead of promoters. Greater significance was found using random genomic regions, confirming the *a priori* assumption that TFs are more likely to bind sequences near genes. This is due to differences in sequence composition between genic regions and non-genic regions,

including an overall higher GC content in coding regions (164). However, with ChIP-chip promoter based foreground sets the relative enrichment of targeted TFs compared to other TFs searched for was usually not more enriched, indicating a higher number of false positives. We therefore suggest, as has been for the similar enrichment of GO terms (91; 92; 88; 89), that for identifying over-represented TFBSs in a foreground versus background set of sequences, that both foreground and background sequences have similar properties, *e.g.* GC content, CpG content, and genic or genomic basis.

7.2 Wet Lab Identification and Analysis of Transcription Factor Binding Sites: Past, Present, and Future

Traditional methods, like the TransFactor kit, luciferase assays, and deletion constructs, only identified one TFBS at a time. Chromatin immunoprecipitation (ChIP) permitted the isolation of all sequences bound in the cell by a given TF, but was also, at first, limited to only analyzing a small number of targets at a time by site specific PCRs. With the invention of the microarray these ChIP fragments (ChIP-chip) could be analyzed on a genomic scale, though still limited by cost and target region (*i.e.* probes on the array). Lately, the costs for genome-wide ChIP analysis has gone down and nowadays the targetting of specific regions can be avoided by the use of ChIP in conjunction with next-generation sequencing machines (ChIP-seq). ChIP-seq also requires less input material and, potentially, can identify low affinity TFBSs (35). However, ChIP-seq is still costly and requires days of preparation. With newer technologies being introduced that permit single molecule sequencing the costs and man hours to produce data will continue to decrease. Already, articles are published using the first of these machines: the Heliscope Single Molecule Sequencer by Helicos (165; 166).

Besides using ChIP-seq and ChIP-chip for direct targeting of TFBSs, other techniques can be used to infer TFBSs. As addressed in chapters 2 and 3, groups of genes with differential expression from expression studies can be used to mine for common sequence patterns indicating shared regulatory elements (*i.e.* TFs). Like ChIP-seq and ChIP-chip, expression applications previously done on a target by target scale have been upgraded to genome wide analysis. Expression, originally analyzed by simple PCRs and gels or qPCR, can be done genome wide with a microarray or in conjunction with next-generation sequencing. Next-generation sequencing for expression analysis proves more precise, reproducible, and sensitive compared to microarrays, likely due to avoiding the background issues of hybridization techniques (36). This also provides data on genes that have similar regulation, for which the regulators (*e.g.* TFs), can be searched for. Still, the location of regulatory regions of such genes has to be determined, such as promoter regions which are often loosely defined as sequences flanking the first exon of a transcript. Techniques like DeepCAGE (Cap-analysis of Gene Expression with high throughput sequencing) can refine this. With these two applications, provided in multiple cell types and conditions, we will have greater quantity and quality of regions to search for TFBSs with *in silico* analysis.

The analysis methods of data from for next-generation sequencing applications, including ChIP-seq, have moved dramatically forward. Initially the primary limitation

was time. Though sufficient for machines with longer reads, for millions of small reads the traditional programs of BLAST (37) or BLAT (38) were not sufficient. The Eland alignment tool provided by Illumina increased speed dramatically, but was limited to a short read length of 32bp. Other programs were introduced that could handle longer reads at high quality, like Rmap (40) and Maq (39), though at a longer run time. Old algorithms were reexamined and the Burrows-Wheeler algorithm took on new life with the current short read alignment standards: Bowtie (42) and BWA (43). Alignments are getting faster and faster, though accuracy should be maintained. The ability to align massive quantities of data will continue to be an important issue as current platforms produce more data per run and future platforms are introduced that provide even larger quantities of sequence.

As outlined in chapter 5, we find binding of regulatory proteins with discrete peaks (most frequently around the transcription start site (TSS)), binding across a gene with a bias for the TSS and transcript end (a so-called "U" shape), and binding across the whole gene (Figure 5.1C-E). Possibly, these patterns are related to the different ways in which p300 and CBP are able to regulate transcription: the local peaks might be associated with genes where p300/CBP bind specifically to the TFs that regulate gene expression in contrast to the gene-wide binding where p300/CBP regulate the expression via histone acetylation to open up the chromatin structure facilitating transcription activation. The combination of association with TFs and histones may account for the "U" shaped binding.

A major problem in ChIP-seq analysis is defining a proper peak, which indicates the target of TF binding in a sequence. These multiple binding patterns could prove troublesome for some current peak detection algorithms. Multiple programs have been designed for ChIP-seq analysis, SISSRs (96), QuEST (97), a pipeline by Kharchenko and colleagues (98), and FindPeaks (99). All of these models are based on strand biases, in which a double peak is created due to the fact that only the 5' ends of all DNA fragments are sequenced and for both strands (Figure 7.2A). When ChIP-seq sample p300 (T30-2, chapter 5) is reanalyzed with GAPSS_B (chapter 4) and we keep tags strand-separated we do see this configuration for many TFBSs (Figure 7.2B). However, this same data also shows cases of very close tags that may not follow these models perfectly (Figure 7.2C). We also see large genomic regions of binding (Figure 5.1D), which we speculate as histone interactions. Since these regions do not represent peaks, but broad binding it is not clear if algorithms will detect these properly. Distinct binding patterns, such as in Figure 7.2B, should perform well with current prediction systems, but further evaluation should be made on broader binding, such as in Figure 5.1D. Future programs should focus on better addressing and identifying multiple peak-shapes, and not just one shape.

The future of ChIP-seq data production and analysis will be faster runs at higher quality, generating more accurate data. Single molecule sequencing allows one to sequence the DNA of a single cell. This will enable more detailed experiments with precise expression and ChIP analysis from a single cell, not a mix of cells common to many cultures and samples.

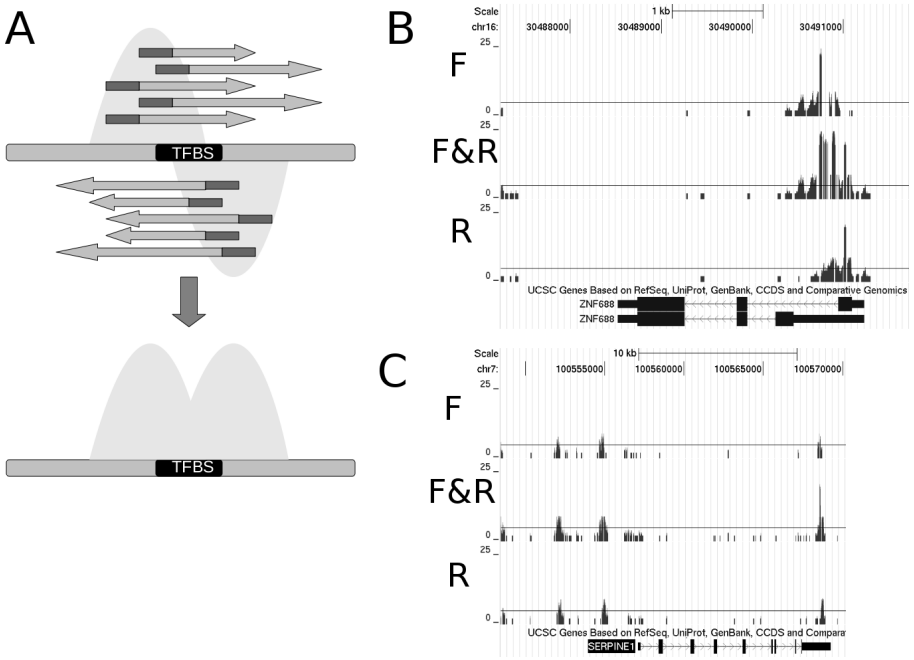


Figure 7.2: Peaks: ChIP-fragments (in full represented by arrows in the top panel A), with darker ends representing what is sequenced are aligned to the genome with strand specificity. The coverage of the sequences are represented by the light "peaks," shown as strand specific (top panel A) and represented as a consensus (bottom panel A). The assumption of sequence tag distribution in programs like FindPeaks, QuEST, and SISSRS is a double peak pattern (A). UCSC browser wiggle tracks (B/C) of tags plotted on the forward (F), reverse (R), or both (F&R) for a p300 ChIP-seq dataset (T30-2 from chapter 5). When we plot p300 ChIP-seq data we see these strand specific patterns for some regions (*e.g. ZNF688*) (B), but for other binding events the strand bias is much less prominent (*e.g. SERPINE1*) (C).

7.3 TFs and Disease

Many diseases are some how associated with function and dysfunction of TFs. This work focuses on TFs involved with two biological processes, myogenesis and cell cycle (control), both of which are linked to multiple diseases (26; 18; 122; 123). In chapters 2 and 3 we identify many TFBSs over-represented in MyoD or Myogenin bound DNA fragments, indicating other TFs that also regulate these fragments. As shown in tables 3.2 and 3.3, these additional TFs already have evidence linking them to the process of myogenesis. If not already identified as disease related, these serve as ideal candidate genes for any disease study involving muscle development and regulation, of which MyoD and Myogenin are master regulators. In chapter 5 we use the same approach to identify TFBSs over-represented in p300 and CBP bound DNA fragments, serving as ideal candidate genes for CBP/p300 related diseases such as Rubinstein-Taybi Syndrome and cancer. Several of these, such as YY1, already have known relations

to cell cycle regulation.

As more ChIP-seq data series become available, we will design better PWMs to search for TFBSs. This knowledge, coupled with whole genome sequencing of patients, will lead to discovering the cause of many diseases. Besides looking at the obvious for mutations (the coding regions of these genes), mutated target TFBS should also be observed. In addition, besides looking for the loss of a TFBS, gain of TFBSs should also be identified. A gain or loss of regulation can lead to dis-regulation and disease. In the future it will be crucial to screen for this on a patient by patient basis, termed personal medicine.

7.4 The Future of Genomics

The development of several platforms that can sequence billions of base-pairs of DNA sequence in less than a week offers new solutions to existing problems, but also generate new problems. We can now look at DNA sequences genome wide without the biases of hybridization that were part of the micro-array era. This technology has moved us closer to having the means to sequence any individual's genome at a reasonable price and speed, a step needed to truly provide personalized medication. However, billions of base-pairs in terabytes of data provides new difficulties in data analysis and storage. The bottleneck in such experiments has dramatically shifted and will continue to shift more from the wet-lab work towards bioinformatic analysis.

These large quantities of data will place demands on storage, access, and interpretation. For storage of the current next-generation sequencer data the NCBI Sequence Read Archive (SRA, *http://www.ncbi.nlm.nih.gov/Traces/sra* (135) has been created. However, we can expect that instead of a lab generating gigabytes of (processed) data in a week, a lab will generate terabytes in an hour. It will be constant competition between increase in data generation and the rapidly decreasing cost of storage. Besides the actual storage devices, the access speed to these devices must be considered. Even if we can keep up with the data storage, analysis will be dramatically slowed down if connections to these storages are not increased. Finally, more methods must be developed to make the data transparent. As next-generation sequencing becomes everyday practice in research groups and clinics, the tools to extrapolate the information must be made intuitive to the average biologist and clinician.

One observation from current next-generation sequencing applications, is that many unannotated parts of the genome have functional support. Therefore the 'old' term "junk" DNA should be appended, or even removed altogether, in the future. There is also functional support for this from other databases (Figure 6.4A), but the challenge is incorporating all of this into a usable and visible means. Evidence also exists that a majority of bases in the genome are transcribed (167). Traditionally DNA that did not code for a protein was termed junk DNA. However, it quickly became apparent that gene regulation occurred in such junk DNA, such as TFBSs. Also some TFs need space between their TFBSs and the gene promoter, so the spacer DNA can not be truly called junk. In addition, there are genes, such as regulatory non-coding RNAs, that encode an RNA, but not a protein. As we discover more and more about the genome it becomes apparent that most nucleotides serve a purpose and thus the term "junk" DNA has become meaningless.

Another fundamental process that should be addressed concerns evidence that

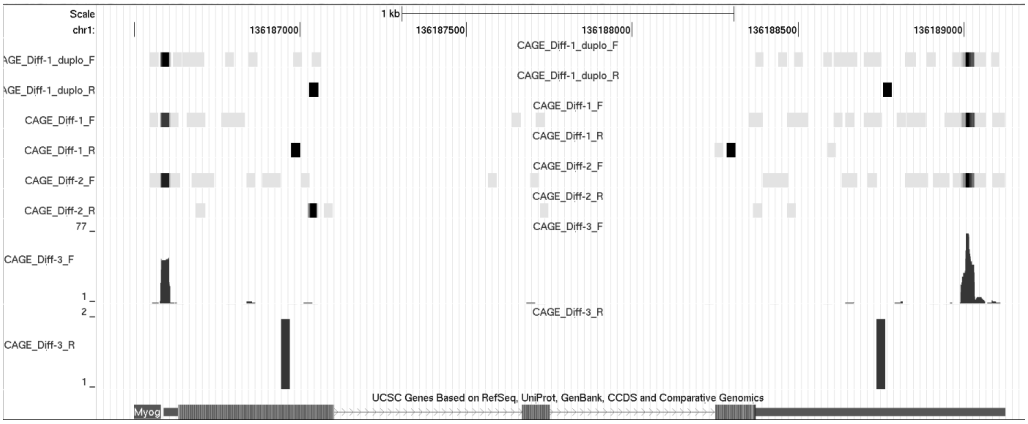


Figure 7.3: Variations from traditional transcription: CAGE tags from all differentiating samples in chapter 6 aligning to both the sense (F) and antisense (R) strands of the myogenin gene. Also, reads align to both 3′ and 5′ ends. Sample 3 is shown as a full wiggle track while other samples are shown as dense tracks to save space. Sample 3 y-axis is indicative of the number of tags at each position, with a maximum per sample of the highest number of tags on a position in this viewing window.

transcription of genes may not always occur on the traditional sense strand. More and more we find evidence for what is termed "antisense" transcription, where a gene is transcribed from the non-coding strand. Antisense transcripts have been previously reported, and in some cases even have higher expression levels than their sense counterparts (168; 36). Though not as prominent as sense strand results, in chapter 5 we do find CAGE and SAGE tags aligning to the non-coding strand. 't Hoen *et al.* (36) have published evidence of SAGE antisense transcription, which even indicated that in 11% of their genes antisense transcription was even more prominent than sense transcription. A human genome-wide study also presented approximately 1600 transcripts with evidence of transcription from both strands (169). We also find binding on the opposite end of the genes as expected (3′ for CAGE and 5′ for SAGE, CAGE example in Figure 6.1). This is also observed for CBP/p300 binding, where in addition to binding transcript starts there is transcript end binding (Figure 5.1A-B). Finding TF binding and CAGE tags at the transcript ends has also been found in other studies (138; 160). In addition, we see in the CAGE/SAGE tags aligning to the gene end on the anti-sense strand. Examples of CAGE tags aligning to the antisense strand, gene end, and both are shown in Figure 7.3. These CAGE/SAGE/ChIP-seq data that show unexpected results (antisense, gene end, or both) were regarded as artifacts in the past, but as many different methods point towards these phenomenons they must be considered as a true biological process. As the increased quantity of less biased genomic data arises the process of reverse and/or antisense transcription will hopefully become more evident.

Next-generation sequencing has provided us with the means to also take the genomics field to the next level. It is becoming more cost-effective and accurate to measure gene expression and genome-wide TF binding. These methods are becoming

more quantitative, with genes expressed in exact numbers of sequences, as apposed to previous methods (*i.e.* probe intensities on an array). In addition, as shown in chapter six, we have the methods to better annotate gene structure, such as TSSs, used in specific cells/tissues under exact conditions. This chapter demonstrates that the current genome annotation, however impressive, is not complete. Though not addressed in this thesis, there are also next-generation applications to identify DNA methylation and chromatin accessibility (via ChIP-seq). The challenge of the bioinformatician in the near future will be combining existing and upcoming information about gene expression, gene structure, DNA methylation, chromatin composition, TF binding, and additional genome properties to construct a more complete model of the entire biological processes in the genome of an animal, including man. This can even be made more complete by combining more fields, such as proteomics, to construct a complete picture of life and its regulation. Besides a picture of life in general, these high throughput methods are, and will increasingly, be used to identify variation with a single individual or population, leading to true personalized medicine and increased effectiveness of health care.

Chapter 8

Summary

Transcription factors (TFs) are an essential part of gene regulation. Mutations in TFs, and their binding sites (TFBSs), can result in muscle diseases such as myotonic dystrophy, rhabdomyosarcoma, Waardenburg syndrome type 2, congenital myasthenia, and diseases related to muscle regeneration, (overview in Martin 2003 (26)). 50% of tumors, resulting from loss of cell cycle control, carry a mutation in the TF p53 (18). Therefore, to cure these diseases and many more it is essential we better understand TFs, their genomic targets, and function.

One research area that has improved greatly with modern sequencing technologies is the study of TFs. Chromatin-immunoprecipitation (ChIP) provides a means to isolate stretches of DNA bound by a protein, such as TFs. These fragments of DNA can be sequenced, resulting in genome wide identification of TFBSs. Previously polymerase-chain reactions (PCRs) and micro-array technologies mostly only looked at target regions, such as promoters, missing many binding regions. Currently, next-generation sequencing of ChIP DNA provides full genome (target free) results, at a lower cost, higher reproducibility, and with the ability to detect low-affinity binding sites.

We have focused on two biological processes of interest: the cell cycle and muscle differentiation. Both are essential processes: cell cycle for replication/division and myogenesis for muscle development and regeneration. Defects in the cell cycle result in death and cancer, whereas muscular dystrophies may result from impaired myogenesis. TFs, such as p300 or CBP for cell cycle control, or MyoD or Myogenin for myogenesis, have been identified to regulate their parent processes, though full details of their binding locations and regulation have not been eluted. Still, many other TFs have not been discovered or related to either process. We have made an approach to further broaden our knowledge of cell cycle and myogenic control through known TFs.

In chapter 2 we present a web application called CORE_TF (Conserved and Over-REpresented Transcription Factor binding sites) that identifies TFs that occur more often in an experimental set of sequences compared to a random set of sequences. It also has the ability to identify TFBSs that are conserved across different organisms. Initially this was developed to identify TFs that potentially regulate co-expressed genes from micro-array studies. However, CORE_TF can also be used with next-generation sequencing expression studies and to identify co-regulators from

micro-array or next-generation sequencing of ChIP samples.

We expanded on CORE_TF's principle of identifying over-represented TFBSs in Chapter 3. Instead of using CORE_TF's Match to identify binding sites, we used a novel tool called Sunflower. Sunflower models competition between TFs for the same nucleotide sequences. This is closer to the actual biological state. After identifying potential TFBSs with Sunflower, we used the same statistical test as CORE_TF to identify TFBSs that are enriched in an experimental set compared to a random set. This process is not as user friendly or fast as CORE_TF, but gives improved results.

As we began to implement our own wet-lab work with a next-generation sequencing platform (Illumina's Genome Analyzer) we realized we needed a general pipeline to begin analysis of our data. Therefore, we developed GAPSS (General Analysis Pipeline for Second-generation Sequencers). As discussed in chapter 4, GAPSS gives us the possibility to quickly edit for contaminating linker sequences, align our data to the genome, make it viewable in a genome browser, and present this data as defined regions.

In chapter 5 we investigated cell cycle control, as regulated by the TFs CBP and p300. ChIP-seq was performed in a model cell line. The data was analyzed initially with the GAPSS pipeline described in chapter 4. By using CORE_TF (chapter 2) we managed to identify TFs that work as partners with CBP and p300. Though these TFs are highly similar and seem to regulate similar genes, we were able to identify targets specific to each TF and potential regulatory partners (e.g. AP-1, AP-2, SP1, and SRF).

Our work in chapters 2 and 3 often relied on analyzing promoter regions. However, often alternative (or previously unannotated) promoters are used during particular processes, in different tissues, and at distinct time points. In chapter 6 we used CAGE and SAGE techniques coupled with next-generation sequencing to provide a better look into the promoters and genes that differed between proliferating and differentiating mouse myoblasts. This used the GAPSS pipeline of chapter 4 for initial data analysis. To prove these novel promoter regions were muscle specific we searched for and found over-representation of muscle specific TFs.

This thesis demonstrates techniques to identify TFs regulating a process, both with novel *in silico* and modern wet lab techniques, such as next-generation sequencing of ChIP DNA. We elucidated the roles of myogenic and cell cycle control TFs, specifically MyoD, Myogenin, CBP, and p300, but these techniques could be applied to transcriptional control of any other biological process.

Chapter 9

Samenvatting

Transcriptiefactoren zijn een essentieel deel van genregulatie. Mutaties in transcriptiefactoren en hun bindingsplaatsen kunnen resulteren in spierziekten zoals myotone dystrofie, rhabdomyosarcoma, Waardenburg syndrome type 2, congenitale myasthenia en andere spierafbraak gerelateerde ziektes (overzicht in Martin 2003 (26)). Vijftig procent van de tumoren die het gevolg zijn van het verliezen van de controle over de celcyclus bevatten een mutatie in transcriptiefactor p53 (18). Om deze ziektes te genezen is het noodzakelijk dat we begrijpen wat transcriptiefactoren zijn, waar deze binden en wat hun functie is.

Een onderzoeksgebied dat enorm verbeterd is met de moderne sequencing technologieën is het bestuderen van transcriptiefactoren. Chromatine-ImmunoPrecipitatie (ChIP) maakt het mogelijk om DNA strengen te isoleren die gebonden zijn aan een eiwit, bijvoorbeeld een transcriptiefactor. Deze DNA fragmenten kunnen gesequenced worden wat resulteert in een genoomwijde identificatie van transcriptiefactor-bindingsplaatsen. Voorheen zijn met polymerase-kettingreactie (PCRs) en micro-arrays vooral specifieke regio's bestudeerd, zoals promotoren, waardoor veel bindingsplaatsen werden gemist. Vandaag de dag zorgt de next-generation sequencing van ChIP DNA voor genoomwijde resultaten tegen lagere kosten en met hogere reproduceerbaarheid. Daarnaast geeft het de mogelijkheid om bindingsplaatsen met een lage affiniteit te detecteren.

We hebben onze aandacht gericht op twee essentiële biologische processen: de celcyclus en spier differentiatie. De celcyclus is belangrijk voor de replicatie en deling, en myogenese voor spierontwikkeling en reparatie. Defecten in de celcyclus resulteren in kanker, en spierdystrofieën zijn het gevolg van nadelig beïnvloede myogenese. Transcriptiefactoren zoals p300 of CBP die zorgen voor controle over de celcyclus, en MyoD en Myogenin voor myogenese, zijn bekend deze processen te reguleren. Volledige details over hun bindingsplaatsen en regulatie zijn echter onbekend. Daarnaast zijn vele andere transcriptiefactoren nog niet ontdekt of gerelateerd aan beide processen. We hebben een aanzet gemaakt om onze kennis over de controle van celcyclus en myogenese via transcriptiefactoren te verbreden.

In hoofdstuk 2 introduceren we de webapplicatie CORE_TF (Conserved and Over-REpresented Transcription Factor binding sites) die transcriptiefactoren

identificeert die vaker voorkomen in een experimentele set sequenties, ten opzichte van een willekeurige set. Het geeft ook de mogelijkheid om transcriptiefactor-bindingsplaatsen te identificeren die geconserveerd zijn tussen verschillende organismen. In eerste instantie is dit ontwikkeld om transcriptiefactoren te vinden die mogelijk co-gexpresseerde genen reguleren binnen micro-array studies. Echter, CORE_TF kan ook worden gebruikt met next-generation sequencing expressie studies en om co-regulatoren te vinden met behulp van data van micro-arrays of next-generation sequencing van ChIP monsters.

Het principe van identificatie van overgerepresenteerde transcriptiefactor-bindingsplaatsen met CORE_TF wordt verder uitgelicht in hoofdstuk 3. In plaats van CORE_TF's Match om bindingsrichtpunten te identificeren, is gebruikt gemaakt van de opkomende software Sunflower. Sunflower modelleert de competiviteit tussen transcriptiefactoren die binden aan dezelfde nucleotide sequenties waarmee dit beter bij de biologische werkelijkheid aansluit. Nadat potentiële bindingsplaatsen zijn gevonden door Sunflower, is dezelfde statistische test gebruikt als in CORE_TF, om transcriptiefactor-bindingsplaatsen te identificeren die verrijkt zijn in een experimentele set ten opzichte van een willekeurige set. Dit proces is minder gebruiksvriendelijk en snel dan CORE_TF, maar geeft betere resultaten.

Op het moment dat ons eigen laboratoriumwerk met een next-generation sequencing machine (Illumina's Genome Analyzer) tot stand kwam, realiseerden wij ons dat een algemeen protocol nodig was om met de data analyse te beginnen. Hiervoor is GAPSS (General Analysis Pipeline for Second-generation Sequencers) ontwikkeld. In hoofdstuk 4 staat beschreven hoe GAPSS de mogelijkheid biedt om snel vervuilende linker sequenties weg te filteren, data naar het genoom te mappen, te visualiseren in een genoom browser en te presenteren in gedefinieerde regio's.

In hoofdstuk 5 wordt de celcyclus controle, gereguleerd door de transcriptiefactoren CBP en p300, nader onderzocht. Hiervoor is een ChIP-seq uitgevoerd op een cellijn model. De data is in eerste instantie geanalyseerd met GAPSS, zoals beschreven in hoofdstuk 4. Met behulp van CORE_TF (hoofdstuk 2) werden transcriptiefactoren geïdentificeerd die samenwerken met CBP en p300. Ondanks dat deze transcriptiefactoren sterk overeenkomen en dezelfde genen lijken te reguleren, waren we alsnog in staat om bindingsplaatsen specifiek voor iedere transcriptiefactor en potentiële regulerende partners te identificeren (bijv. AP-1, AP-2, SP1 en SRF).

Het werk in hoofdstuk 2 en 3 berust vooral op het analyseren van promotoren. Niettemin worden alternatieve (of voorheen onbekende) promotoren gebruikt tijdens bepaalde processen, in verschillende weefsels en in verschillende tijdspunten. In hoofdstuk 6 worden CAGE en SAGE technieken gekoppeld aan next-generation sequencing om een beter inzicht te krijgen in promotoren en genen die verschillen tussen vermenigvuldigende en differentiërende muisspierstamcellen. Hierbij is voor de initiële data analyse gebruik gemaakt van GAPSS uit hoofdstuk 4. Om te bewijzen dat deze nieuwe promotoren specifiek zijn voor spieren is er met succes gezocht naar overrepresentatie van spierspecifieke transcriptiefactoren.

Dit proefschrift demonstreert technieken om proces regulerende transcriptiefactoren te identificeren door middel van nieuwe, in silico en moderne laboratorium technieken zoals next-generation sequencing van ChIP DNA.

Daarnaast verduidelijken we de rol van myogenese en celcyclus controlerende transcriptiefactoren. In het bijzonder MyoD, Myog, CBP en p300. Deze technieken kunnen echter toegepast worden op de transcriptionele controle van elk willekeurig biologisch proces.

Abbreviations

Commonly used Abbreviations in this thesis:

bp = base pair(s)
CAGE = Cap Analysis of Gene Expression
ChIP = chromatin immuniprecipitation
ChIP-(on-)chip = ChIP hybridized to a microarray
ChIP-PET = ChIP with paired-end ditag sequencing
ChIP-seq = ChIP with next-generation sequencing
DeepCAGE = next-generation sequencing of CAGE sequences
DeepSAGE = next-generation sequencing of SAGE sequences
HAT = histone acetyltransferase
Kb = kilo base(s)
NGS = next-generation sequencers or sequencing
PCR = polymerase chain reaction
PWM = position weight matrix
qPCR = quantitative (real-time) PCR
RT-PCR = reverse transcriptase PCR
SAGE = Serial Analysis of Gene Expression
TF = transcription factor
TFBS = transcription factor binding site
TSS = transcription start site
UTR = untranslated region

The four nucleobases of DNA:

A = adenine
T = thymine
C = cytosine
G = guanine

Bibliography

[1] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de laBastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K.,

Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., deJong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y. J. (2001) *Nature* **409(6822)**, 860–921.

[2] Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley,

K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., and Lander, E. S. (2002) *Nature* **420(6915)**, 520–62.

[3] Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009) *Nucleic Acids Res* **37(Database issue)**, D690–7.

[4] Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004) *Genome Res* **14(1)**, 160–9.

[5] Lannigan, D. A. and Notides, A. C. (1989) *Proc Natl Acad Sci U S A* **86(3)**, 863–7.

[6] Lee, J. S., Lee, C. H., and Chung, J. H. (1998) *Proc Natl Acad Sci U S A* **95(3)**, 969–74.

[7] Reed, B. D., Charos, A. E., Szekely, A. M., Weissman, S. M., and Snyder, M. (2008) *PLoS Genet* **4(7)**, e1000133.

[8] Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engstrom, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006) *Nat Genet* **38(6)**, 626–35.

[9] Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005) *Nature* **436(7052)**, 876–80.

[10] Gardiner-Garden, M. and Frommer, M. (1987) *J Mol Biol* **196(2)**, 261–82.

[11] Wei, C.-L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J., Zhao, X. D., Chew, J.-L., Lee, Y. L., Kuznetsov, V. A., Sung, W.-K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H.-H., and Ruan, Y. (2006) *Cell* **124(1)**, 207–19.

[12] Arney, K. L. and Fisher, A. G. (2004) *J Cell Sci* **117(Pt 19)**, 4355–63.

[13] Miele, A. and Dekker, J. (2008) *Mol Biosyst* **4(11)**, 1046–57.

[14] King, M. C. and Wilson, A. C. (1975) *Science* **188(4184)**, 107–16.

[15] Satyanarayana, A. and Kaldis, P. (2009) *Oncogene* **28(33)**, 2925–39.

[16] Malumbres, M. and Barbacid, M. (2009) *Nat Rev Cancer* **9(3)**, 153–66.

[17] Leake, R. (1996) *Ann N Y Acad Sci* **784**, 252–62.

[18] Polager, S. and Ginsberg, D. (2009) *Nat Rev Cancer* **9(10)**, 738–48.

[19] Grossman, S. R. (2001) *Eur J Biochem* **268(10)**, 2773–8.

[20] Yao, T. P., Oh, S. P., Fuchs, M., Zhou, N. D., Ch'ng, L. E., Newsome, D., Bronson, R. T., Li, E., Livingston, D. M., and Eckner, R. (1998) *Cell* **93(3)**, 361–72.

[21] Tanaka, Y., Naruse, I., Hongo, T., Xu, M., Nakahata, T., Maekawa, T., and Ishii, S. (2000) *Mech Dev* **95(1-2)**, 133–45.

[22] Charge, S. B. P. and Rudnicki, M. A. (2004) *Physiol Rev* **84(1)**, 209–38.

[23] Asakura, A. and Rudnicki, M. (2002) *In: Mouse Development* , 253278.

[24] Pownall, M. E., Gustafsson, M. K., and Emerson, C. P. J. (2002) *Annu Rev Cell Dev Biol* **18**, 747–83.

[25] Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y., and Dynlacht, B. D. (2005) *Genes Dev* **19(5)**, 553–69.

[26] Martin, P. T. (2003) *Curr Opin Pharmacol* **3(3)**, 300–8.

[27] Chomczynski, P. and Sacchi, N. (2006) *Nat Protoc* **1(2)**, 581–5.

[28] Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) *Science* **270(5235)**, 484–7.

[29] Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003) *Proc Natl Acad Sci U S A* **100(26)**, 15776–81.

[30] Nielsen, K. L., Hogh, A. L., and Emmersen, J. (2006) *Nucleic Acids Res* **34(19)**, e133.

[31] Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T. T., Tang, M.-H. E., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A., and Carninci, P. (2009) *Genome Res* **19(2)**, 255–65.

[32] Harbers, M. and Carninci, P. (2005) *Nat Methods* **2(7)**, 495–502.

[33] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009) *Nat Methods* **6(5)**, 377–82.

[34] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000) *Science* **290(5500)**, 2306–9.

[35] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007) *Nat Methods* **4(8)**, 651–7.

[36] 'tHoen, P. A. C., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H. A. M., deMenezes, R. X., Boer, J. M., vanOmmen, G.-J. B., and denDunnen, J. T. (2008) *Nucleic Acids Res* **36(21)**, e141.

[37] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J Mol Biol* **215(3)**, 403–10.

[38] Kent, W. J. (2002) *Genome Res* **12(4)**, 656–64.

[39] Li, H., Ruan, J., and Durbin, R. (2008) *Genome Res* **18(11)**, 1851–8.

[40] Smith, A. D., Xuan, Z., and Zhang, M. Q. (2008) *BMC Bioinformatics* **9**, 128.

[41] Schatz, M. C. (2009) *Bioinformatics* **25(11)**, 1363–9.

[42] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) *Genome Biol* **10(3)**, R25.

[43] Li, H. and Durbin, R. (2009) *Bioinformatics* **25(14)**, 1754–60.

[44] Zerbino, D. R. and Birney, E. (2008) *Genome Res* **18(5)**, 821–9.

[45] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006) *Nucleic Acids Res* **34(Web Server issue)**, W369–73.

[46] Bailey, T. L. and Elkan, C. (1994) *Proc Int Conf Intell Syst Mol Biol* **2**, 28–36.

[47] Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993) *Science* **262(5131)**, 208–14.

[48] Neuwald, A. F., Liu, J. S., and Lawrence, C. E. (1995) *Protein Sci* **4(8)**, 1618–32.

[49] Pavesi, G., Mauri, G., and Pesole, G. (2004) *Brief Bioinform* **5(3)**, 217–36.

[50] Stormo, G. D. (1990) *Methods Enzymol* **183**, 211–21.

[51] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003) *Nucleic Acids Res* **31(1)**, 374–8.

[52] TRANSFAC website: *www.biobase-international.com/cgi-bin/biobase/transfac/start.cgi.*

[53] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004) *Nucleic Acids Res* **32(Database issue)**, D91–4.

[54] JASPAR website:
http://jaspar.genereg.net/.

[55] Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003) *Nucleic Acids Res* **31(13)**, 3576–9.

[56] Hoffman, M. M. and Birney, E. (2010) *Genome Res* **20(5)**, 685–92.

[57] Elkon, R., Linhart, C., Sharan, R., Shamir, R., and Shiloh, Y. (2003) *Genome Res* **13(5)**, 773–80.

[58] Gumucio, D. L., Shelton, D. A., Zhu, W., Millinoff, D., Gray, T., Bock, J. H., Slightom, J. L., and Goodman, M. (1996) *Mol Phylogenet Evol* **5(1)**, 18–32.

[59] Hardison, R. C., Oeltjen, J., and Miller, W. (1997) *Genome Res* **7(10)**, 959–66.

[60] Sui, S. J. H., Fulton, D. L., Arenillas, D. J., Kwon, A. T., and Wasserman, W. W. (2007) *Nucleic Acids Res* **35(Web Server issue)**, W245–52.

[61] Hooghe, B., Hulpiau, P., vanRoy, F., and Bleser, P. D. (2008) *Nucleic Acids Res* **36(Web Server issue)**, W128–32.

[62] Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007) *Nucleic Acids Res* **35(Database issue)**, D610–7.

[63] sorttable.js: *http://www.kryogenix.org/code/browser/sorttable/.*

[64] Kobes, R. to access the cephes math library, by Moshier, SL. math::cephes perl interface:
http://search.cpan.org/dist/Math-Cephes/lib/Math/Cephes.pod.

[65] Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003) *Genome Res* **13(1)**, 103–7.

[66] Cao, Y., Kumar, R. M., Penn, B. H., Berkes, C. A., Kooperberg, C., Boyer, L. A., Young, R. A., and Tapscott, S. J. (2006) *EMBO J* **25(3)**, 502–11.

[67] Smyth, G., Yang, Y., and Speed, T. (2003) *Methods Mol Biol* **224**, 111–136.

[68] Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W. (2005) Bioinformatics and Computational Biology Solutions using R and Bioconductor., Springer, New York.

[69] Alibes, A., Yankilevich, P., Canada, A., and Diaz-Uriarte, R. (2007) *BMC Bioinformatics* **8**, 9.

[70] Benjamini, Y. and Hochberg, Y. (1995) *J R Statist Soc B* **57**, 289–300.

[71] Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., and Birney, E. (2005) *Nucleic Acids Res* **33(Database issue)**, D447–53.

[72] Edmondson, D. G., Brennan, T. J., and Olson, E. N. (1991) *J Biol Chem* **266(32)**, 21343–6.

[73] Banerjee-Basu, S. and Buonanno, A. (1993) *Mol Cell Biol* **13(11)**, 7019–28.

[74] Brunetti, A. and Goldfine, I. D. (1990) *J Biol Chem* **265(11)**, 5960–3.

[75] Wyzykowski, J. C., Winata, T. I., Mitin, N., Taparowsky, E. J., and Konieczny, S. F. (2002) *Mol Cell Biol* **22(17)**, 6199–208.

[76] Hestand, M. S., vanGalen, M., Villerius, M. P., vanOmmen, G.-J. B., denDunnen, J. T., and 'tHoen, P. A. C. (2008) *BMC Bioinformatics* **9**, 495.

[77] Tokovenko, B., Golda, R., Protas, O., Obolenskaya, M., and El'skaya, A. (2009) *Nucleic Acids Res* **37(7)**, e49.

[78] Zambelli, F., Pesole, G., and Pavesi, G. (2009) *Nucleic Acids Res* **37(Web Server issue)**, W247–52.

[79] Marstrand, T. T., Frellsen, J., Moltke, I., Thiim, M., Valen, E., Retelska, D., and Krogh, A. (2008) *PLoS One* **3(2)**, e1623.

[80] Zheng, J., Wu, J., and Sun, Z. (2003) *Nucleic Acids Res* **31(7)**, 1995–2005.

[81] Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008) *Nature* **451(7178)**, 535–40.

[82] Sinha, S. (2006) *Bioinformatics* **22(14)**, e454–63.

[83] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1999) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, UK.

[84] Roider, H. G., Manke, T., O'Keeffe, S., Vingron, M., and Haas, S. A. (2009) *Bioinformatics* **25(4)**, 435–42.

[85] Roider, H. G., Lenhard, B., Kanhere, A., Haas, S. A., and Vingron, M. (2009) *Nucleic Acids Res* **37(19)**, 6305–15.

[86] Ng, P. and Keich, U. (2008) *Genome Inform* **21**, 15–26.

[87] Ng, P. and Keich, U. (2008) *Bioinformatics* **24(19)**, 2256–7.

[88] Khatri, P. and Draghici, S. (2005) *Bioinformatics* **21(18)**, 3587–95.

[89] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009) *Nucleic Acids Res* **37(1)**, 1–13.

[90] Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C. J., Jenster, G., and Kors, J. A. (2008) *Genome Biol* **9(6)**, R96.

[91] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) *Nat Genet* **25(1)**, 25–9.

[92] Camon, E., Barrell, D., Lee, V., Dimmer, E., and Apweiler, R. (2004) *In Silico Biol* **4(1)**, 5–6.

[93] Ramos, Y. F., Hestand, M. S., Verlaan, M., Krabbendam, E., Ariyurek, Y., vanGalen, M., vanDam, H., vanOmmen, G. J., denDunnen, J. T., Zantema, A., and 'tHoen, P. A. (2010) *Nucleic Acids Res* **Epub**, 2010 Apr 30.

[94] Rice, P., Longden, I., and Bleasby, A. (2000) *Trends Genet* **16(6)**, 276–7.

[95] Slater, G. S. C. and Birney, E. (2005) *BMC Bioinformatics* **6**, 31.

[96] Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008) *Nucleic Acids Res* **36(16)**, 5221–31.

[97] Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008) *Nat Methods* **5(9)**, 829–34.

[98] Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008) *Nat Biotechnol* **26(12)**, 1351–9.

[99] Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. M. (2008) *Bioinformatics* **24(15)**, 1729–30.

[100] Jongeneel, C. V., Iseli, C., Stevenson, B. J., Riggins, G. J., Lal, A., Mackay, A., Harris, R. A., O'Hare, M. J., Neville, A. M., Simpson, A. J. G., and Strausberg, R. L. (2003) *Proc Natl Acad Sci U S A* **100(8)**, 4702–5.

[101] Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008) *Nat Methods* **5(12)**, 1005–10.

[102] Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008) *Nucleic Acids Res* **36(16)**, e105.

[103] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002) *Genome Res* **12(6)**, 996–1006.

[104] Farnham, P. J. (2009) *Nat Rev Genet* **10(9)**, 605–16.

[105] Moore, M. J. and Proudfoot, N. J. (2009) *Cell* **136(4)**, 688–700.

[106] Sonenberg, N. and Hinnebusch, A. G. (2009) *Cell* **136(4)**, 731–45.

[107] Lu, T.-Y., Kao, C.-F., Lin, C.-T., Huang, D.-Y., Chiu, C.-Y., Huang, Y.-S., and Wu, H.-C. (2009) *J Cell Biochem* **108(1)**, 315–25.

[108] Selaru, F. M., David, S., Meltzer, S. J., and Hamilton, J. P. (2009) *Am J Gastroenterol* **104(8)**, 1910–2.

[109] Szyf, M. (2009) *Clin Rev Allergy Immunol* , Epub ahead of print.

[110] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007) *Cell* **129(4)**, 823–37.

[111] Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q., and Zhao, K. (2008) *Nat Genet* **40(7)**, 897–903.

[112] Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007) *Nature* **448(7153)**, 553–60.

[113] Welboren, W.-J., vanDriel, M. A., Janssen-Megens, E. M., vanHeeringen, S. J., Sweep, F. C., Span, P. N., and Stunnenberg, H. G. (2009) *EMBO J* **28(10)**, 1418–28.

[114] Smeenk, L., vanHeeringen, S. J., Koeppel, M., vanDriel, M. A., Bartels, S. J. J., Akkers, R. C., Denissov, S., Stunnenberg, H. G., and Lohrum, M. (2008) *Nucleic Acids Res* **36(11)**, 3639–54.

[115] Wederell, E. D., Bilenky, M., Cullum, R., Thiessen, N., Dagpinar, M., Delaney, A., Varhol, R., Zhao, Y., Zeng, T., Bernier, B., Ingham, M., Hirst, M., Robertson, G., Marra, M. A., Jones, S., and Hoodless, P. A. (2008) *Nucleic Acids Res* **36(14)**, 4549–64.

[116] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007) *Science* **316(5830)**, 1497–502.

[117] Wang, Z., Zang, C., Cui, K., Schones, D. E., Barski, A., Peng, W., and Zhao, K. (2009) *Cell* **138(5)**, 1019–31.

[118] Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009) *Nature* **457(7231)**, 854–8.

[119] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Calcar, S. V., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007) *Nat Genet* **39(3)**, 311–8.

[120] Kalkhoven, E. (2004) *Biochem Pharmacol* **68(6)**, 1145–55.

[121] Goodman, R. H. and Smolik, S. (2000) *Genes Dev* **14(13)**, 1553–77.

[122] Roelfsema, J. H. and Peters, D. J. M. (2007) *Expert Rev Mol Med* **9(23)**, 1–16.

[123] Iyer, N. G., Ozdag, H., and Caldas, C. (2004) *Oncogene* **23(24)**, 4225–31.

[124] Chrivia, J. C., Kwok, R. P., Lamb, N., Hagiwara, M., Montminy, M. R., and Goodman, R. H. (1993) *Nature* **365(6449)**, 855–9.

[125] Duyndam, M. C., vanDam, H., Smits, P. H., Verlaan, M., van derEb, A. J., and Zantema, A. (1999) *Oncogene* **18(14)**, 2311–21.

[126] Puri, P. L., Avantaggiati, M. L., Balsano, C., Sang, N., Graessmann, A., Giordano, A., and Levrero, M. (1997) *EMBO J* **16(2)**, 369–83.

[127] Partanen, A., Motoyama, J., and Hui, C. C. (1999) *Int J Dev Biol* **43(6)**, 487–94.

[128] Kung, A. L., Rebel, V. I., Bronson, R. T., Ch'ng, L. E., Sieff, C. A., Livingston, D. M., and Yao, T. P. (2000) *Genes Dev* **14(3)**, 272–7.

[129] Shikama, N., Lutz, W., Kretzschmar, R., Sauter, N., Roth, J.-F., Marino, S., Wittwer, J., Scheidweiler, A., and Eckner, R. (2003) *EMBO J* **22(19)**, 5175–85.

[130] Bordoli, L., Husser, S., Luthi, U., Netsch, M., Osmani, H., and Eckner, R. (2001) *Nucleic Acids Res* **29(21)**, 4462–71.

[131] Tullai, J. W., Schaffer, M. E., Mullenbrock, S., Sholder, G., Kasif, S., and Cooper, G. M. (2007) *J Biol Chem* **282(33)**, 23981–95.

[132] Denissov, S., vanDriel, M., Voit, R., Hekkelman, M., Hulsen, T., Hernandez, N., Grummt, I., Wehrens, R., and Stunnenberg, H. (2007) *EMBO J* **26(4)**, 944–54.

[133] Freeman, M. F. and Tukey, J. W. (1950) *Ann Math Statist* **21(4)**, 607–611.

[134] Dennis, G. J., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003) *Genome Biol* **4(5)**, P3.

[135] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008) *Nucleic Acids Res* **36(Database issue)**, D13–21.

[136] Deng, L., de laFuente, C., Fu, P., Wang, L., Donnelly, R., Wade, J. D., Lambert, P., Li, H., Lee, C. G., and Kashanchi, F. (2000) *Virology* **277(2)**, 278–95.

[137] Chow, C.-W. and Davis, R. J. (2006) *Cell* **127(5)**, 887–890.

[138] Gilchrist, D. A., Fargo, D. C., and Adelman, K. (2009) *Methods* **48(4)**, 398–408.

[139] Cho, H., Orphanides, G., Sun, X., Yang, X. J., Ogryzko, V., Lees, E., Nakatani, Y., and Reinberg, D. (1998) *Mol Cell Biol* **18(9)**, 5355–63.

[140] Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engstrom, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., and Wahlestedt, C. (2005) *Science* **309(5740)**, 1564–6.

[141] Gordon, S., Akopyan, G., Garban, H., and Bonavida, B. (2006) *Oncogene* **25(8)**, 1125–42.

[142] Seto, E., Lewis, B., and Shenk, T. (1993) *Nature* **365(6445)**, 462–4.

[143] Zhou, Q., Gedrich, R. W., and Engel, D. A. (1995) *J Virol* **69(7)**, 4323–30.

[144] Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008) *Biotechniques* **45(1)**, 81–94.

[145] Sterrenburg, E., Turk, R., 'tHoen, P. A. C., vanDeutekom, J. C. T., Boer, J. M., vanOmmen, G.-J. B., and denDunnen, J. T. (2004) *Neuromuscul Disord* **14(8-9)**, 507–18.

[146] Turk, R., Sterrenburg, E., van derWees, C. G. C., deMeijer, E. J., deMenezes, R. X., Groh, S., Campbell, K. P., Noguchi, S., vanOmmen, G. J. B., denDunnen, J. T., and 'tHoen, P. A. C. (2006) *FASEB J* **20(1)**, 127–9.

[147] Jelier, R., 'tHoen, P. A. C., Sterrenburg, E., denDunnen, J. T., vanOmmen, G.-J. B., Kors, J. A., and Mons, B. (2008) *BMC Bioinformatics* **9**, 291.

[148] Tomczak, K. K., Marinescu, V. D., Ramoni, M. F., Sanoudou, D., Montanaro, F., Han, M., Kunkel, L. M., Kohane, I. S., and Beggs, A. H. (2004) *FASEB J* **18(2)**, 403–5.

[149] Sartorelli, V. and Caretti, G. (2005) *Curr Opin Genet Dev* **15(5)**, 528–35.

[150] Vencio, R. Z. N., Brentani, H., Patrao, D. F. C., and Pereira, C. A. B. (2004) *BMC Bioinformatics* **5**, 119.

[151] Huber, W., vonHeydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002) *Bioinformatics* **18 Suppl 1**, S96–104.

[152] Cohen, C. D., Klingenhoff, A., Boucherot, A., Nitsche, A., Henger, A., Brunner, B., Schmid, H., Merkle, M., Saleem, M. A., Koller, K.-P., Werner, T., Grone, H.-J., Nelson, P. J., and Kretzler, M. (2006) *Proc Natl Acad Sci U S A* **103(15)**, 5682–7.

[153] Dohr, S., Klingenhoff, A., Maier, H., deAngelis, M. H., Werner, T., and Schneider, R. (2005) *Nucleic Acids Res* **33(3)**, 864–72.

[154] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., and Edgar, R. (2009) *Nucleic Acids Res* **37(Database issue)**, D885–90.

[155] Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., Cook, B. P., Dufault, M. R., Ferguson, A. T., Gao, Y., He, T. C., Hermeking, H., Hiraldo, S. K., Hwang, P. M., Lopez, M. A., Luderer, H. F., Mathews, B., Petroziello, J. M., Polyak, K., Zawel, L., and Kinzler, K. W. (1999) *Nat Genet* **23(4)**, 387–8.

[156] Buskin, J. N. and Hauschka, S. D. (1989) *Mol Cell Biol* **9(6)**, 2627–40.

[157] Olson, E. N. (1990) *Genes Dev* **4(9)**, 1454–61.

[158] Nishida, W., Nakamura, M., Mori, S., Takahashi, M., Ohkawa, Y., Tadokoro, S., Yoshida, K., Hiwada, K., Hayashi, K., and Sobue, K. (2002) *J Biol Chem* **277(9)**, 7308–17.

[159] Suzuki, H., Forrest, A. R. R., vanNimwegen, E., Daub, C. O., Balwierz, P. J., Irvine, K. M., Lassmann, T., Ravasi, T., Hasegawa, Y., deHoon, M. J. L., Katayama, S., Schroder, K., Carninci, P., Tomaru, Y., Kanamori-Katayama, M., Kubosaki, A., Akalin, A., Ando, Y., Arner, E., Asada, M., Asahara, H., Bailey, T., Bajic, V. B., Bauer, D., Beckhouse, A. G., Bertin, N., Bjorkegren, J., Brombacher, F., Bulger, E., Chalk, A. M., Chiba, J., Cloonan, N., Dawe, A., Dostie, J., Engstrom, P. G., Essack, M., Faulkner, G. J., Fink, J. L., Fredman, D., Fujimori, K., Furuno, M., Gojobori, T., Gough, J., Grimmond, S. M., Gustafsson, M., Hashimoto, M., Hashimoto, T., Hatakeyama, M., Heinzel, S., Hide, W., Hofmann, O., Hornquist, M., Huminiecki, L., Ikeo, K., Imamoto, N., Inoue, S., Inoue, Y., Ishihara, R., Iwayanagi, T., Jacobsen, A., Kaur, M., Kawaji, H., Kerr, M. C., Kimura, R., Kimura, S., Kimura, Y., Kitano, H., Koga, H., Kojima, T., Kondo, S., Konno, T., Krogh, A., Kruger, A., Kumar, A., Lenhard, B., Lennartsson, A., Lindow, M., Lizio, M., Macpherson, C., Maeda, N., Maher, C. A., Maqungo, M., Mar, J., Matigian, N. A., Matsuda, H., Mattick, J. S., Meier, S., Miyamoto, S., Miyamoto-Sato, E., Nakabayashi, K., Nakachi, Y., Nakano, M., Nygaard, S., Okayama, T., Okazaki, Y., Okuda-Yabukami, H., Orlando, V., Otomo, J., Pachkov, M., Petrovsky, N., Plessy, C., Quackenbush, J., Radovanovic, A., Rehli, M., Saito, R., Sandelin, A., Schmeier, S., Schonbach, C., Schwartz, A. S., Semple, C. A., Sera, M., Severin, J., Shirahige, K., Simons, C., Laurent, G. S., Suzuki, M., Suzuki, T., Sweet, M. J., Taft, R. J., Takeda, S., Takenaka, Y., Tan, K., Taylor, M. S., Teasdale, R. D., Tegner, J., Teichmann, S., Valen, E., Wahlestedt, C., Waki, K., Waterhouse, A., Wells, C. A., Winther,

O., Wu, L., Yamaguchi, K., Yanagawa, H., Yasuda, J., Zavolan, M., Hume, D. A., Arakawa, T., Fukuda, S., Imamura, K., Kai, C., Kaiho, A., Kawashima, T., Kawazu, C., Kitazume, Y., Kojima, M., Miura, H., Murakami, K., Murata, M., Ninomiya, N., Nishiyori, H., Noma, S., Ogawa, C., Sano, T., Simon, C., Tagami, M., Takahashi, Y., Kawai, J., and Hayashizaki, Y. (2009) *Nat Genet* **41(5)**, 553–62.

[160] Fejes-Toth, Sotirova, Sachidanandam, Assaf, Hannon, Kapranov, Foissac, Willingham, Duttagupta, Dumais, and Gingeras (2009) *Nature* **457(7232)**, 1028–32.

[161] Sabourin, L. A. and Rudnicki, M. A. (2000) *Clin Genet* **57(1)**, 16–25.

[162] Funk, W. D. and Wright, W. E. (1992) *Proc Natl Acad Sci U S A* **89(20)**, 9484–8.

[163] Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004) *Curr Opin Struct Biol* **14(3)**, 283–91.

[164] Guigo, R. and Fickett, J. W. (1995) *J Mol Biol* **253(1)**, 51–60.

[165] Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009) *Nature* **461(7265)**, 814–8.

[166] Pushkarev, D., Neff, N. F., and Quake, S. R. (2009) *Nat Biotechnol* **27(9)**, 847–52.

[167] Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C.,

Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Calcar, S. V., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., deBakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and deJong, P. J. (2007) *Nature* **447(7146)**, 799–816.

[168] Allo, M., Buggiano, V., Fededa, J. P., Petrillo, E., Schor, I., de laMata, M., Agirre, E., Plass, M., Eyras, E., Elela, S. A., Klinck, R., Chabot, B., and Kornblihtt, A. R. (2009) *Nat Struct Mol Biol* **16(7)**, 717–24.

[169] Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K., and Rotman, G. (2003) *Nat Biotechnol* **21(4)**, 379–86.

List of Publications

-Ramos YFM*, Hestand MS*, Verlaan M, Krabbendam E, Ariyurek Y, van Galen M, van Dam H, van Ommen GJ, den Dunnen JT, Zantema A*, 't Hoen PA*. Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. Nucleic Acids Res. 2010 Apr 30. [Epub ahead of print]
*=equal first or last authorship

-Hestand MS, van Galen M, Villerius MP, van Ommen GJ, den Dunnen JT, 't Hoen PA. 2008. CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. BMC Bioinformatics. Nov 26;9(1):495.

-Wiersma AC, Millon LV, Hestand MS, Van Oost BA, Bannasch DL. 2005. Canine COL4A3 and COL4A4: sequencing, mapping and genomic organization. DNA Seq. Aug;16(4):241-51.

-van den Berg L, Kwant L, Hestand MS, van Oost BA, Leegwater PA. 2005. Structure and variation of three canine genes involved in serotonin binding and transport: the serotonin receptor 1A gene (htr1A), serotonin receptor 2A gene (htr2A), and serotonin transporter gene (slc6A4). J Hered. 96(7):786-96.

Curriculum Vitae

Matthew Hestand was born January 3, 1979, in Lexington, KY, USA. He graduated from Tates Creek High School in Lexington, KY, in 1997. He then did a Bachelors of Science in Biology and Bachelors of Arts in Philosophy at the University of Kentucky, USA, graduating in December 2001. He was also a graduate of the university's Honors Program. During this time he worked and did research in the university's traditional equine blood-typing lab, including a mouse/alpaca syntenic mapping project. After his bachelors he worked as a lab technician in a virology lab at the Livestock Disease Diagnostic Center in Lexington, Kentucky.

From Fall 2002 until September 2004 he did a Masters in Chemistry with the Genomics and Bioinformatics program at the University of Utrecht, NL. This included several projects: sequencing and characterizing the canine serotonin transporter gene *SLC6A4*, (in silico) comparative genomics of the *MURR1* gene, and a masters thesis comparing EBI (Ensembl) and NCBI gene annotation and visualization.

In April 2005 he began a Center for Biomedical Genetics sponsored PhD at Leiden University, situated in the department of Human and Clinical Genetics at the Leiden University Medical Center under the supervision of Peter-Bram 't Hoen, Johan den Dunnen, and Gert-Jan van Ommen. The PhD focused on transcriptional regulation, with the results presented in this thesis. During this time he won the best application showcase at the NBIC (Netherlands BioInformatics Centre) Conference 2009 for his web-based application CORE_TF. Also, during and since this time he has been actively involved in education with supervision of several short and one long-term student projects and teaching and coordination of several next-generation sequencing courses.

Starting April 2009 he became the coordinator for next-generation sequencing bioinformatics at the Leiden Genome Technology Center in Leiden, NL as part of the NBIC BioAssist program.